Title
A unified mechanism for intron and exon definition and back-splicing.

Journal
Nature, 573(7774)

ISSN
0028-0836

Authors
Li, Xueni
Liu, Shiheng
Zhang, Lingdi
et al.

Publication Date
2019-09-01

DOI
10.1038/s41586-019-1523-6

Peer reviewed

Published in final edited form as:

Nature. 2019 September ; 573(7774): 375–380. doi:10.1038/s41586-019-1523-6.

A unified mechanism for intron and exon definition and back-splicing

Xueni Li[1,†], **Shiheng Liu**[2,3,†], **Lingdi Zhang**[1], **Aaron Issaian**[1], **Ryan C. Hill**[1], **Sara Espinosa**[1], **Shasha Shi**[1], **Yanxiang Cui**[3], **Kalli Kappel**[4], **Rhiju Das**[4,5], **Kirk C. Hansen**[1], **Z. Hong Zhou**[2,3,*], **Rui Zhao**[1,6,*]

[1.]Department of Biochemistry and Molecular Genetics, University of Colorado Denver Anschutz Medical Campus, Aurora, CO 80045, USA

[2.]Department of Microbiology, Immunology, and Molecular Genetics, UCLA, Los Angeles, CA 90095, USA

[3.]Electron Imaging Center for Nanomachines University of California, Los Angeles (UCLA), Los Angeles, CA 90095, USA

[4.]Biophysics Program, Stanford University, Stanford, CA USA

[5.]Department of Biochemistry and Department of Physics, Stanford University, Stanford, CA USA

[6.]RNA Bioscience Initiative, School of Medicine, University of Colorado Denver Anschutz Medical Campus, Aurora, CO 80045, USA

Summary

The molecular mechanisms of exon definition and back-splicing are fundamental unanswered questions in pre-mRNA splicing. Here we report cryoEM structures of the yeast E complex assembled on introns, providing the first view of the earliest event in the splicing cycle that commits pre-mRNAs to splicing. The E complex architecture suggests that the same spliceosome can assemble across an exon, which either remodels to span an intron for canonical linear splicing (typically on short exons) or catalyzes back-splicing generating circRNA (on long exons). The model is supported by our experiments demonstrating that E complex assembled on the yeast EFM5 or HMRA1 middle exon can be chased into circRNA when the exon is sufficiently long.

[*]Correspondence and requests for materials should be addressed to rui.zhao@cuanschutz.edu or hong.zhou@ucla.edu".

Author Contributions

R.Z. and Z.H.Z. conceived the project; X.L. prepared and optimized the sample; X.L., L.Z., S.E., and S.S. performed biochemical analyses; S.L. and Y.C. recorded and processed the EM data; A.I., R.H. and K.C.H. performed mass spectrometry analyses; S.L. built the atomic models; K.K. and R.D. built partial U1 snRNA model and the minimal exon model in the A complex; R.Z., S.L., X.L., and Z.H.Z. analyzed and interpreted the models; S.L., X.L. and R.Z. prepared the illustrations; R.Z., S.L. and Z.H.Z. wrote the paper; and all authors contributed to the editing of the manuscript.

[†]These authors contributed equally to this work and are listed alphabetically.

The authors declare no competing interests.

Data Availability

The coordinate files have been deposited in the Protein Data Bank (6N7P for the Ubc4 complex and 6N7R for the Act1 complex). The cryoEM maps have been deposited in the Electron Microscopy Data Bank (EMD-0360 for the Ubc4 complex and EMD-0361 for the Act1 complex).

This simple model unifies intron definition, exon definition, and back-splicing through the same spliceosome in all eukaryotes and should inspire experiments in many other systems to understand the mechanism and regulation of these processes.

The spliceosome forms sequentially the E, A, Pre-B, B, Bact, B*, C, C*, P, and ILS complexes through the splicing cycle. CryoEM structures of all but one *S. cerevisiae* (yeast) spliceosomal complexes [1,2] provided valuable information on later stages of the splicing cycle. There is, however, a lack of structural and mechanistic understanding of the E complex formation, the earliest event that initiates the splicing cycle. Thus, how the splicing machinery accurately defines introns and exons remains a fundamental unanswered question. In yeast which typically contain small introns and large exons, intron definition, where the spliceosome initially recognizes and assembles across an intron, seems to dominate [3]. On the other hand, exon definition [4] prevails in vertebrate, where small exons and large introns are prevalent. In the exon definition model, the spliceosome recognizes and assembles across an exon first. However, in order to splice out introns, it was assumed that the exon definition complex (EDC) needs to be remodeled to a cross-intron complex. Support for the exon definition model is largely circumstantial, and biochemical and structural analyses of the exon definition process are limited. Although the EDC seems to be similar to the intron definition complex (IDC) in composition [5,6], we do not know if the two complexes differ in their structural organization and how an EDC remodels to span an intron.

In addition to canonical splicing, a peculiar back-splicing reaction generates a class of circular RNAs (circRNAs) observed in diverse eukaryotic species, prompting the speculation that back-splicing is also an ancient and conserved feature of eukaryotic gene expression pathway [7]. CircRNAs play roles in the regulation of their host genes or microRNAs, aging, and other disease processes [8]. Although canonical splicing signals and spliceosome are needed for circRNA production [9], the exact players and mechanism of back-splicing remain unknown.

To fill these gaps, we set out to obtain molecular details of the earliest step of the yeast splicing cycle that commits a pre-mRNA to splicing. In yeast, intron recognition is initiated by the binding of U1 snRNP on the 5’ splice site (ss) [10-12], and the recognition of the branch point sequence (BPS) by the BBP and Mud2 heterodimer (the 3’ ss is not recognized until much later), forming the E complex (also referred to as the CC2 complex) [13]. Here we report the cryoEM structure of the yeast E complex assembled on either the Act1 pre-mRNA or the Ubc4 pre-mRNA. These structures and subsequent biochemical analyses reveal a unified mechanism for intron definition, exon definition and remodeling, and back-splicing-mediated circRNA biogenesis.

In vitro assembled E complex is functional

After discovering that E complex purified from yeast is too heterogenous for structural determination, we assembled the E complex *in vitro* using uncapped 3xMS2-Act1 pre-mRNA (M3-Act1) and purified U1 snRNP, BBP and Mud2 proteins (referred to as the Act1 complex). The complex was purified sequentially by the MS2 tag and the CBP tag on U1A

and Mud2. After RNase H cleavage of M3-Act1 into two fragments (Extended Data Fig. 1a-b), the MS2 tag still pulled down all U1 snRNP proteins, BBP, and Mud2 (Fig. 1a), confirming that U1 snRNP and BBP/Mud2 interact instead of being simply tethered through M3-Act1. In addition, the assembled Act1 complex can be chased into spliced M3-Act1 in U1 snRNA depleted yeast extract (Fig. 1b, lane 5). Although excess Act1-3xMS2 pre-mRNA (Act1-M3) can effectively compete with free M3-Act1 for splicing (Fig. 1b, lane 2), it cannot compete with assembled Act1 complex (Fig. 1b, lane 6). These data indicate that our assembled E complex has not fallen apart significantly in the splicing extract and is functional.

5’ ss recognition is facilitated by proteins and RNA secondary structures

We determined the cryoEM structure of the Act1 complex to 3.2 Å resolution (Extended Data Fig. 2-4, Extended Data Table 1). After observing low resolutions in several key areas, we also assembled the E complex on a capped Ubc4 pre-mRNA, crosslinked the complex with BS3, and determined its structure to 3.6 Å resolution. The overall structures of the two complexes are similar (Extended Data Fig. 5a) and subsequent discussions refer to their common features unless otherwise stated.

In these structures, the 5’ ss basepairs with the 5’ end of U1 snRNA (Fig. 2a-b), which is stabilized by U1C and Luc7 proteins, similar to that observed in the yeast A and pre-B structures [14,15]. In addition, a homology model of the yeast nuclear cap binding protein (NCBP) complex can be fitted as a rigid body into the density upstream of nucleotide −9 of Ubc4 (Fig. 2c), likely binding to the pre-mRNA cap. The RRM domain of U1-70K is shifted toward NCBP in the Ubc4 complex compared to the uncapped Act1 complex (Extended Data Fig. 5a), suggesting that NCBP directly interacts with U1-70K RRM and providing a possible mechanism by which NCBP recruits U1 snRNP and facilitates splicing of cap-proximal introns [16,17]. In both complexes, the RRM2 domain of Nam8 is positioned to bind to the intronic region immediately downstream of nt +13 (Fig. 2c), illustrating the structural basis of Nam8’s role in facilitating 5’ ss recognition [18].

A striking feature in the Act1 complex is a ~25bp double helix on a binding surface formed by many positively charged residues on the C-terminal tail of Prp39 and the N-terminal domain of Prp42, as well as the C-terminal domain of U1C (Fig. 2d). Such double helix density is also observed in the Pre-B complex structure and is tentatively modeled as part of U2 snRNA [15]. Our Act1 complex is *in vitro* assembled and contains no U2 snRNA (Extended Data Fig. 5b). Furthermore, no such double helix exists in the Ubc4 complex, supporting that this helix is part of the Act1 pre-mRNA. Although we were unable to model specific nucleotides, a weak density connects this helix to the 5’ ss, suggesting that it belongs to the region downstream of the 5’ ss. The 5’ ss to BPS region (265nt) of the Act1 intron is predicted to form long stem-like structures, while the same region in Ubc4 (58nt) contains a much shorter possible secondary structure (Extended Data Fig. 6a), potentially explaining why a stem-like structure is observed in the Act1 but not the Ubc4 complex. Mutation of this region in the Act1-Cup1 reporter [19] that abolishes extensive secondary structures (Extended Data Fig. 6b) leads to significant pre-mRNA accumulation compared to the WT (Fig. 2e), suggesting that this secondary structure facilitates splicing. Our structures

of the E and P complexes [20] therefore provided direct evidence that the intronic regions of pre-mRNA can form highly ordered secondary structures, which may help bring key intronic elements close together and whose direct interaction with proteins may also facilitate spliceosomal assembly.

The 5’ ss and BPS are bridged by intrinsically flexible Prp40

A critical event in the first step of the splicing cycle is to define the intron by bringing together the 5’ ss and BPS in which U1 snRNP protein Prp40 seems to be a key player [13]. Prp40 contains two N-terminal WW domains, a ~60-residue linker, and six C-terminal FF domains. In the region between U1-70K and Luc7 in the Ubc4 complex structure, there is a boomerang-shaped density that matches well with the crystal structures of tandem FF domains connected by long helices [21,22] (Fig. 2f, Extended Data Fig. 4g). (This density is not obvious in the Act1 complex, possibly because the Act1 complex is not cross-linked with BS3.) There is weak density connecting the boomerang-shaped density and U1-70K (Fig. 2f), and the C-terminal FF domains crosslinks to U1-70K in the Ubc4 complex (while the N-terminal and middle FF domains crosslink to Luc7 and Snu71) (Extended Data Fig. 7a). These observations led us to assign the boomerang-shaped density as the Prp40 FF4-6 domains (although we cannot rule out the possibility of this density being the other tandem FF domains such as FF3-5), which is also consistent with our observation that Prp40 FF1-3 domain interacts with Luc7 (Extended Data Fig. 7b).

Prior biochemical analyses have shown that Prp40 forms a stable dimer with Snu71 and a trimer with Snu71-Luc7 [23-25], and the Prp40 WW domains directly interact with the N-terminal domain of BBP [13,26]. BBP also forms non-exclusive interactions with both Prp40 and Mud2 [13], and BBP directly binds to the BPS of pre-mRNA [27]. In the Act1 complex structure, there is a large volume of weak density close to the pre-mRNA double helix (Extended Data Fig. 4j). The density can be best interpreted as the BBP/Mud2 dimer for three reasons: Its location corresponds roughly to where U2 SnRNP is in the A complex structure [14] (Extended Data Fig. 4j); crosslinking and mass spectrometry analyses indicate that BBP/Mud2 is located in this region (Extended Data Fig. 7a); BBP/Mud2 are the only proteins left in the E complex that are large enough to fill the volume of this density. This density is not obvious in the Ubc4 complex structure, potentially because Ubc4 lacks the pre-mRNA helix that brings the BPS close to the 5’ ss. Prp40 therefore bridges both ends of the intron by interacting with U1-70K and Snu71-Luc7 of U1 snRNP through its FF domains, and interacting with BBP through its WW domains. The entire ~60-residue linker region between the WW and FF domains is predicted to be disordered (Extended Data Fig. 7c), explaining why density corresponding to BBP/Mud2 is difficult to observe.

Exon definition occurs in yeast

The E complex architecture, in particular the relative positions between the 5’ ss and the BPS where BBP binds, suggests that the same E complex can form across an exon. Instead of connecting the upstream 5’ ss to a downstream BPS through an intron (Fig. 3a), the BPS can be connected to a downstream 5’ ss through an exon (Fig. 3b). Similarly, the A complex structure [14] suggests that the same A complex could also form across exons (Fig. 3b).

Modeling using Rosetta RNP-denovo method [28] suggests that only 28nt between the upstream BP and downstream 5' ss is needed to span the U2 snRNP and U1 snRNP in the A complex (Extended Data Fig. 8a). The minimal distance connecting the same BP and 5' ss is likely similar or smaller in the E complex, given the similar spatial position and smaller size of BBP/Mud2 compared to U2 snRNP (Extended Data Fig. 4j). On the other hand, adding the tri-snRNP to form the pre-B complex forces an ~30° increase in the angle between U1 snRNP and U2 SF3b [14,15] (Fig. 3b). A relatively short exon may hinder this conformational change and also create steric hindrance for the addition of the bulky tri-snRNP (Fig. 3b). This may signal to the spliceosome that an EDC has formed and provide an opportunity for the upstream 5' ss to interact with the tri-snRNP to form an intron-spanning B complex (Fig. 3b). We use EDC to refer to spliceosomal complexes assembled across an exon, which can be an exon-defined E, A, or unstable pre-B complex.

To test whether the E complex can form across a yeast exon, we truncated the multi-intronic *DYN2* gene to only contain its middle exon and partial flanking introns (Dyn2 IEI, Extended Data Fig. 8b). Spliceosomal complexes assemble on either Dyn2 WT or IEI pre-mRNAs (using the same protocol as the Act1 complex) contain the same protein components in similar quantities, even after RNase H cleavage between the BPS and 5' ss (Extended Data Fig. 8c-e). Furthermore, 2D classifications of negative-stain images of the Dyn2 IEI complex resemble those of the Act1 and Ubc4 complexes (Extended Data Fig. 8f). These observations support the formation of E complex across the Dyn2 middle exon *in vitro*.

We next asked whether exon-definition occurs *in vivo* in yeast, by evaluating whether mutation of splice sites bordering the Dyn2 middle exon negatively affect splicing of both flanking introns, a hallmark used to support the exon definition model in vertebrates [4]. We generated a BPS mutation in intron 1 (I1-BP mutant), a 5' ss mutation in intron 2 (I2-5'ss mutant), and a double mutation on *DYN2* gene (Extended Data Fig. 8b). We demonstrated that the I2-5' ss and I1-BP mutations impaired the splicing of intron 1 and intron 2, respectively (Fig. 4a). We further evaluated the splicing products of WT and each mutant using PCR and primers located in exons 1 and 3 (Fig. 4b). If Dyn2 splicing is solely governed by intron definition, we would observe retention of the intron where these mutations reside (with minimal effect on the distal intron), generating products containing a single intron (255 and 271bp bands). On the other hand, if Dyn2 splicing is solely governed by exon definition, the mutations would lead to the retention of both introns (*i.e.*, accumulation of the 351bp pre-mRNA band) or exon skipping (the 152bp band), but not any product containing a single intron (indicating that the distal intron was successfully spliced). The fact that we observed both pre-mRNA accumulation and single-intron-containing products (Fig. 4b, lanes 4 and 5) suggest that both intron definition and exon definition contribute to Dyn2 splicing *in vivo*. We observed exon skipping for the I1-BP mutant but not the I2-5' ss mutant, consistent with previous observations [29]. This observation differs from that in the mammalian system where exon definition mutation leads to predominantly exon skipping, likely because intron definition also contributes to Dyn2 splicing, leading to co-transcriptional splicing of intron 1 in I2-5' ss mutant which prevents exon skipping [29]. Taken together, our results demonstrate that exon definition occurs for a fraction of the Dyn2 transcripts *in vivo* in yeast.

The EDC catalyzes back-splicing on long exons

An intriguing prediction of our exon definition model is that if the exon connecting the BP and downstream 5' ss is long enough, it will not create much steric hindrance and will allow tri-snRNP to join the pre-B complex and complete the rest of the splicing cycle (Fig. 3c). As a result, the 5' ss downstream of the exon will be back-spliced to the upstream 3' ss, generating a circRNA through the same transesterification reaction used by canonical splicing (Fig. 3d). Supporting this hypothesis, 7 of the 10 multi-intron genes in *S. cerevisiae* form circRNA products [7].

To test this model, we purified yeast spliceosome using TAP-tagged Cef1 (a strategy used to purify and determine the cryoEM structures of multiple spliceosomal complexes) from the $Prp22^{H606A}$ mutant strain defective in exon release [30]. As expected, purified spliceosomes contained spliced mRNA and lariat for yeast single-intronic gene *RPP1B*, as well as the unique T-branch and circRNA for multi-intronic genes *EFM5* and *HMRA1* (Fig. 5a, Extended Data Fig. 9a). These results establish that Cef1-purified spliceosome contains both canonical and back-splicing products.

Further supporting this model (Fig. 3c-d), we showed using RT-PCR that the EFM5 IEI construct on an expression plasmid generated a RNase R-resistant circRNA corresponding to exon 2 *in vivo* (Fig. 5b, lane 10, Extended Data Fig. 9b). Mutating the BPS or 5' ss or shortening exon 2 to 63 nt abolishes circRNA formation (Fig. 5b, lanes 11, 12, and 14). E complex assembled on *in vitro* transcribed EFM5 IEI-101-M3 (exon 2 shortened to 101nt and 3xMS2 at the 3' end) (Extended Data Fig. 9c) can be chased into circRNA in U1-depleted yeast extract in the presence of excess competing IEI-101 RNA (Fig. 5c). To ensure the generality of our observation, we carried out the same experiments using another yeast multi-intronic gene *HMRA1* and obtained the same conclusion (Extended Data Fig. 9d-e). Taken together, these results support that exon definition occurs in yeast across the EFM5 or HMRA1 middle exon, which catalyzes back-splicing and generates circRNA when this exon is sufficiently long.

Discussion

It was previously unclear whether the EDC is the same or different from the IDC. The architecture of the E complex makes it immediately apparent that the same complex can form across either intron or exon without the need of additional components or structural rearrangement, and the same can be deduced for the A complex. The structures of the E and A complexes predict a minimal BP to 5' ss distance (28nt for the A complex and likely a similar or smaller number for the E complex) in order for exon definition to occur. An exon that is above this minimum but still relatively short potentially makes it difficult for tri-snRNP to join the spliceosome. This leads to the spliceosome to stall at the pre-B stage and fail to handoff the 5' ss from U1 to U6, providing an opportune point for the spliceosome to remodel into an intron-spanning B complex involving the upstream 5' ss. This model is consistent with the observation in mammalian systems where tri-snRNP is loosely associated with the EDC, and only becomes stably associated when a 5'ss-containing RNA oligo is added and the EDC is converted to a B-like intron-spanning complex [6]. Supporting

our exon definition model, we showed that intron definition and exon definition both contribute to yeast Dyn2 splicing *in vivo* (Fig. 4). Although yeast has few multi-intronic genes and exon definition is clearly not the driving force of splicing, our results provide the proof of principle evidence that both intron and exon definition can occur through the same spliceosomal structure in most or all species, even on the same pre-mRNA. Whether intron or exon definition is dominant *in vivo* is likely determined by gene architecture (such as the length of introns or exons) and other factors (such as exonic or intronic enhancers or suppressors and their associated proteins, RNA secondary structures, transcription processivity, and nucleosome positioning, *etc*).

CircRNA generated by back-splicing of exons has attracted increasing attention, but its origin and biogenesis have largely remained a mystery [8]. Although exon definition was speculated to play a role in back-splicing [31,32], it is unclear which of the canonical spliceosomal components are required and what exact signals are being recognized that makes an exon forming circRNA instead of participating in canonical splicing. Our results demonstrate that back-splicing is catalyzed by exon-definition complexes on long exons (or multiple exons) not remodeled to intron-spanning complexes, suggesting that circRNA is a natural byproduct of spliceosome-mediated splicing in all eukaryotic species. This model is consistent with what was envisioned by the Wilusz group based on competition between back-splicing and canonical splicing [31] and with previous observation that long but not short exon (when flanked by the same intronic sequences) can form circRNAs in human cells [33]. Indeed, the average exon length in circRNA is 690 nt [34], much longer than the median length of 120 nt for human exons [35]. The long exon inevitably lowers the efficiency of initial exon-definition, contributing to the low frequency of back-splicing and circRNA production. Intronic complementary sequences flanking the exon and RNA-binding proteins potentially increase the efficiency of initial exon-definition and facilitate circRNA production [8]. These RNA elements or proteins may also bring opposite ends of different exons close together for back-splicing, generating circRNAs containing multiple exons. It is worth noting that, accurately speaking, the distance between the upstream BP and downstream 5’ ss across an exon instead of the exon length in yeast pre-mRNAs determines the fate of the EDC, since the 3’ ss is not recognized in early yeast spliceosomes. However, given the generally short distance between yeast BP and 3’ ss (19 nt for *EFM5* and 10 nt for *HMRA1* intron 1) [36], exon lengths ultimately play a major role in determining the outcome of EDC remodeling.

In summary, our E complex structure enabled us to propose a simple model that unifies intron definition, exon definition, and back-splicing, without needing a different spliceosome for each process. This model is supported by our biochemical analyses performed exclusively in yeast which is now positioned to serve as a well-defined model system to understand exon definition or back-splicing. This model likely holds true for all eukaryotes, although many *cis* or *trans* factors (including RNA, protein, transcription, nucleosome, *etc.*) may act as modulators to promote or suppress a particular process. In vertebrate, most exons are short which is likely the main signal for EDC remodeling, but other factors may facilitate remodeling of EDC assembled on long exons and lower the efficiency of back-splicing. This model should inspire experiments in many other systems to understand the mechanism and regulation of exon definition and back-splicing, some of the most fundamental unanswered questions in pre-mRNA splicing.

Methods

Yeast E complex assembly and purification

The coding regions of yeast BBP and Mud2 were amplified by PCR using genomic *S. cerevisiae* DNA as templates. BBP fused to an N-terminal protein A (protA) tag was inserted between a GPD promoter and a CYC1 terminator, and the resulting expression cassette was cloned into pRS414 to generate the pRS414/GPD-protA-BBP plasmid. Similarly, Mud2 with or without a C-terminal Calmodulin Binding Peptide (CBP) tag was cloned into pRS416 vectors to generate the pRS416/GPD-Mud2-CBP or pRS416/GPD-Mud2 plasmid. Six liters of BCY123 cells harboring both plasmids were grown in -URA-TRP selective media to OD600=3-4. The cells were flash-frozen in liquid nitrogen to form yeast "popcorns" and cryogenically ground using a SPEX 6870 Freezer/Mill. The frozen cell powder was thawed at room temperature and re-suspended in lysis buffer (50 mM Tris-HCl, pH 8.0, 400 mM NaCl, 0.1% NP-40, 1mM DTT) with protease inhibitor cocktails (Sigma-Aldrich) and 1 mM Benzamidine. The cell lysate was first centrifuged at 27,485 × g for 1 hr in a GSA rotor (Sorvall) and the supernatant was further centrifuged at 167,424 × g in a 45Ti rotor (Beckman) for 1.5 hr at 4°C. The supernatant was incubated with 2 ml of IgG Sepharose-6 Fast Flow resin (GE Healthcare) overnight at 4°C. The resin was first washed with IgG washing buffer (20 mM Tris-HCl, pH 8.0, 350 mM NaCl, 0.05% NP-40, 0.5 mM DTT, 1 mM Benzamidine and protease inhibitor cocktails), then with buffer containing 250 mM and 150 mM NaCl. The BBP/Mud2 dimer was released by TEV protease in TEV150 buffer (20 mM Tris-HCl, pH 8.0, 150 mM NaCl, 0.02% NP-40, 0.5 mM DTT).

The Act1 pre-mRNA used in this paper is consisted of a 73 nt 5′ exon, the 302 nt intron that lacks a cryptic branch point sequence, and a 167 nt 3′ exon [37]. The Ubc4 pre-mRNA is consisted of a 20 nt 5′ exon, a 95 nt intron, and a 32 nt 3′ exon [38]. The Dyn2 wildtype pre-mRNA is consisted of three exons (22 nt, 23 nt, and 35 nt in lengths) which are separated by two introns (96 nt and 80 nt in lengths). The Dyn2 Intron-Exon-Intron (IEI) pre-mRNA is consisted of intron 1 without the first 9 nt, the middle exon, and intron 2 truncated right before the branch point sequence. The EFM5 IEI-101 pre-mRNA is consisted of intron 1 without the first 10 nt, the middle exon shortened to 101 nt, and intron 2 truncated to 9 nt upstream of the BPS. The HMRA1 IEI-246 wild type pre-mRNA is consisted of intron 1 without the first 10 nt, the entire middle exon, and intron 2 truncated to 2 nt upstream of the BPS. The HMRA1 IEI-246 pre-mRNA was generated after mutating the underlined part of the last 41 nt of its middle exon from 5'-CAAAGAAATGTGGCATTACTCCACTTCAAGTAAGAGTTTGG-3' to 5'-ACTAATGCCACTACTTTACTCCACTTCAAGTAAGAGTTTGG-3'. This modification enables us to use specific primers to detect only the exogenous but not endogenous HMRA1 in a WT yeast strain. DNA templates for *in vitro* transcription were generated after the addition of three copies of MS2 stem loops to the 5′-end of the *ACT1* gene or to the 3'-end of the *UBC4*, *DYN2, EFM5,* and *HMRA1* genes. Pre-mRNA substrates were generated by run-off transcription from linearized plasmid DNA templates, and capped using Vaccinia Capping System (New England Biolabs) if indicated.

To obtain the yeast complex E for structural studies, the Act1 or Ubc4 pre-mRNA substrate was bound to the MBP-MS2 fusion protein and mixed with purified U1 snRNP [23] and BBP/Mud2 dimer (or BBP/Mud2-CBP in the case of Act1), then applied to amylose resin (New England Biolabs) pre-washed with buffer G120 (20 mM HEPES, pH 7.9, 120 mM KCl, 0.01% NP-40). After 3 h incubation at 4 °C, the resin was washed and eluted with buffer G120 containing 10 mM maltose. Elutions were pooled and applied to 100 μL of calmodulin resin (Agilent) pre-washed with washing buffer (20 mM Hepes, pH7.9, 120 mM KCl, 2 mM CaCl2, 1 mM imidazole, 0.01% NP-40), and incubated for 3 hr at 4 °C. The resin was washed with washing buffer, and eluted 6 times with 100 μL eluting buffer (20 mM Hepes, pH7.9, 120 mM KCl, 2 mM EGTA) each time. The elutions containing the most concentrated E complex were used for cryoEM imaging. Crosslinked sample was prepared by treating the complex with 1 mM BS3 (Thermo Fisher) on ice for 30 min, and subsequently quenched with 50 mM Tris (pH8.0).

CryoEM sample preparation and imaging

For cryoEM sample optimization, an aliquot of 3 μl of sample (~0.2-0.5 μM) was applied onto a glow-discharged lacey carbon film-coated copper grid (400 mesh, Ted Pella). The grid was blotted with Grade 595 filter paper (Ted Pella) and flash-frozen in liquid ethane with a FEI Mark IV Vitrobot. A FEI TF20 cryoEM instrument was used to screen grids. CryoEM grids with optimal particle distribution and ice thickness were obtained by varying the gas source (air using PELCO easiGlow™, target vacuum of 0.37 mbar, target current of 15 mA; or H_2/O_2 using Gatan Model 950 advanced plasma system, target vacuum of 70 mTorr, target power of 50 W) and time for glow discharge, the volume of applied samples, chamber temperature and humidity, blotting time and force, as well as wait time before blotting. Our best grids for the Act1 complex were obtained with 50 s glow discharge using air and with the Vitrobot sample chamber set at 12°C temperature, 100% humidity, 2.5 s blotting time, −3 blotting force and 20 s wait time. The best grids for Ubc4 complex were obtained with 60 s glow discharge using air and with the Vitrobot sample chamber set at 12°C temperature, 100% humidity, 3 s blotting time, 1 blotting force and 60 s wait time.

Optimized cryoEM grids were loaded into a FEI Titan Krios electron microscope with a Gatan Imaging Filter (GIF) Quantum LS device and a post-GIF K2 Summit direct electron detector. The microscope was operated at 300 kV with the GIF energy-filtering slit width set at 20 eV. Movies were acquired with Leginon [39] by electron counting in either super-resolution mode at a pixel size of 0.68 Å/pixel (Act1 complex) or counting mode at a pixel size of 1.36 Å/pixel (Ubc4 complex). A total number of 40 frames were acquired in 8 seconds for each movie, giving a total dose of ~30 e-/Å2/movie.

Drift correction for movie frames

Frames in each movie were aligned for drift correction with the GPU-accelerated program MotionCor2 [40]. The first frame was skipped during drift correction due to concern of more severe drift/charging of this frame. Two averaged micrographs, one with dose weighting and the other without dose weighting, were generated for each movie after drift correction. The averaged micrographs have a calibrated pixel size of 1.36 Å on the specimen scale. The averaged micrographs without dose weighting were used only for defocus determination and

the averaged micrographs with dose weighting were used for all other steps of image processing.

Structure determination for the Act1 *complex*

For the Act1 complex, the defocus value of each averaged micrograph was determined by CTFFIND4 [41] to be ranging from −1.5 to −3 μm. Initially, a total of 3,589,121 particles were automatically picked from 11,283 averaged micrographs without reference using Gautomatch (http://www.mrc-lmb.cam.ac.uk/kzhang). The particles were boxed out in dimensions of 352 × 352 square pixels square and binned to 176 × 176 square pixels (pixel size of 2.72 Å) before further processing by the GPU accelerated RELION2.1. The reported U1 model (EMD-8622) was low-pass filtered to 60 Å to serve as an initial model for 3D classification. After one round of 3D classification, only the classes exhibiting features characteristic of the E complex (*e.g.*, 5'ss and pre-mRNA helix binding to U1 snRNP) were kept, which contained 1,852,842 particles. Several iterations of reference-free 2D classification were subsequently performed to remove "bad" particles (*i.e.*, classes with fuzzy or un-interpretable features), yielding 1,108,069 good particles. Auto-refinement of these particles by RELION yielded a map with an average resolution of 5.44 Å ("Step 1" in Extended Data Fig. 2c).

Next, we performed two rounds of *focused classification* on the pre-mRNA helix region of the E complex to further eliminate those particles without the pre-mRNA helix ("Step2" in Extended Data Fig. 2c). The first round of this focused classification generated one good class containing 390,792 particles. These particles were un-binned to 352 × 352 square pixels (pixel size of 1.36 Å) and subjected to another round of focused classification. We re-centered the particles from one best class and removed duplications based on the unique index of each particle given by RELION.

The 270,587 un-binned, unique particles (7.5% of all particles) resulting from the focused classification were subjected to a final step of 3D auto-refinement ("Step 3" in Extended Data Fig. 2c). The two half maps from this auto-refinement step were subjected to RELION's standard post-processing procedure. The final map of the Act1 complex has an average resolution of 3.2 Å based on RELION's gold-standard FSC (see below).

Structure determination for the Ubc4 complex

For the Ubc4 complex, the defocus value of each averaged micrograph was determined by CTFFIND4 to be ranging from −1.5 to −3 μm. Initially, a total of 1,924,710 particles were automatically picked from 8,997 averaged micrographs without reference using Gautomatch. The particles were boxed out in dimensions of 384 × 384 square pixels square and binned to 192 × 192 square pixels (pixel size of 2.72 Å) before further processing by the GPU accelerated RELION2.1. The reported U1 model (EMD-8622) was low-pass filtered to 60 Å to serve as an initial model for 3D classification. After one round of 3D classification, only the classes showing features corresponding to the E complex (*e.g.*, 5'ss binding to U1 snRNP) were kept, which contained 800,735 particles. Several iterations of reference-free 2D classification were subsequently performed to remove bad particles (i.e., classes with

fuzzy or un-interpretable features), yielding 756,303 good particles ("Step1" in Extended Data Fig. 3c).

Next, we performed another two rounds of 3D classification to further improve the ratio of the intact E complex (*e.g.*, Prp40, NCBP1/NCBP2, and Nam8 binding) ("Step2" in Extended Data Fig. 3c). During each round of the 3D classification, only one class showed features corresponding to the intact E complex monomer (Likely due to the cross-linking reagent used for the Ubc4 complex, one class from the second round of 3D classification exhibiting features characteristic of the E complex dimer). These good particles from the final round of 3D classification were un-binned to 384 × 384 square pixels (pixel size of 1.36 Å). We re-centered these particles and removed duplications based on the unique index of each particle given by RELION.

The resulting 124,825 un-binned, unique particles (6.5% of all particles) were subjected to a final step of 3D auto-refinement ("Step 3" in Extended Data Fig. 3c). The two half maps from this auto-refinement step were subjected to RELION's standard post-processing procedure. The final map of the Ubc4 complex has an average resolution of 3.6 Å based on RELION's gold-standard FSC (see below).

Resolution assessment

All resolutions reported above are based on the "gold-standard" FSC 0.143 criterion [42]. FSC curves were calculated using soft spherical masks and high-resolution noise substitution was used to correct for convolution effects of the masks on the FSC curves [43]. Prior to visualization, all maps were sharpened by applying a negative B-factor which was estimated using automated procedures [44].

Local resolution was estimated using ResMap [45]. The overall quality of the maps for Act1 and Ubc4 complexes is presented in Extended Data Figs. 2b-d and 3b-d, respectively. Data collection and reconstruction statistics are presented in Extended Data Table 1.

Model building and refinement

To aid subunit assignment and model building, we took advantage of the reported U1 structure (PDB code: 5UZ5, 3.7 Å), which were fitted into the Ubc4 complex density map by UCSF CHIMERA [46]. The central regions of the Ubc4 complex have resolutions ranging from 3.0 to 4.5 Å (Extended Data Fig. 3f); thus protein and RNA components in these regions were rebuilt manually using COOT [47]. Briefly, for protein subunits that match well with the densities in the Ubc4 complex structure, we manually adjusted their side chain conformation and, when necessary, moved their main chains to match the density map. For protein subunits that exhibit significant main chain mismatches or have not been identified, we built atomic models *de novo*. To do so, sequence assignment was mainly guided by visible densities of amino acid residues with bulky side chains, such as Trp, Tyr, Phe, and Arg. Other residues including Gly and Pro also helped the assignment process. Unique patterns of sequence segments containing such residues were utilized for validation of residue assignment.

For the RNA region near 5'ss (nt −2 to +8 of pre-mRNA with respect to the exon-intron junction; nt 1-10 of the U1 snRNA), the well-defined nucleotide densities, along with the base pairs between U1 snRNA and pre-mRNA, facilitated the RNA model building process. RNA model building in these regions was performed *de novo* in COOT. For the central regions of U1 snRNA, the previous U1 snRNA model was adjusted for their base conformation and, when necessary, for their main chains to match the density map. The RNA components were subsequently adjusted using RCrane [48] and ERRASER [49].

Models built for the protein and RNA subunits in these central regions include: U1-70K (aa 1-91), U1C (aa 3-197), U1A (aa 2-46, 55-125, 133-148), Prp42 (aa 1-544), Prp39 (aa 288-553, 561-627), Nam8 (aa 291-425, 432-449, 492-523), Snu56 (aa 43-170, 185-295), Snu71 (aa 1-52), Luc7 (aa 4-19, 38-140, 172-244), Sm ring; the core regions of U1 snRNA (Extended Data Fig. 4), pre-mRNA (nt −2 to +8 with respect to the exon-intron junction). The long helix interacting with ZnF2 and the coiled coil domain of Luc7 was traced with poly-alanine, which likely belongs to Snu71 since deletion of the coiled coil domain of Luc7 reduces its interaction with Snu71 and Prp40 but only Snu71 has isolated long helices (Extended Data Fig. 7B).

Resolutions for the periphery of the Ubc4 complex were more varied, ranging from 6 Å to 25 Å, insufficient for *de novo* atomic modeling. The following proteins were built with homology modeling using I-TASSER server and rigidly docked into the low-pass filtered map of the 3.6 Å map using CHIMERA: RRM domain of U1-70K (aa 94-188), N-terminal region of Prp39 (aa 43-285), and RRM2 domain of Nam8 (aa 161-242). The homology model of NCBP1/NCBP2 heterodimer (NCBP1:616 aa 36-861; NCBP2: aa 19-156) was rigidly fitted into its local refinement map. In addition, the periphery region has a boomerang-shaped density which matches well with that of the crystal structures of tandem FF domains connected by long helices [21,22]. We assigned this density as the Prp40 FF4-6 domains, considering that DSSO crosslinking in the Ubc4 complex and mass spectrometry analyses demonstrate that the C-terminal FF domains crosslink to U1-70K (Extended Data Fig. 7A) and that there is weak density connecting the boomerang-shaped density and U1-70K, although we cannot rule out the possibility of this density being the other tandem FF domains such as FF3-5. The FF4 (aa 355-413, from I-TASSER), FF5 (aa 427-488, from I-TASSER) and FF6 (aa 491-552, PDB: 2KFD) domains were rigidly docked into the low-pass filtered map using CHIMERA, and manually connected using COOT with the long helix between FF domains (Fig. 2F).

Except for nucleotides 27-33 at the tip of Stem-Loop 1 and the last three nucleotides 566-568, the entire U1 snRNA is now modeled with DRRAFTER [50]. The estimated mean RMSD accuracies for the DRRAFTER models are: The estimated mean RMSD accuracies for the DRRAFTER models are: 0.4 Å for residues 39-41, 4.3 Å for residues 97-103, 0.7 Å for residues 175-177, 3.5 Å for residues 202-236, 3.0 Å for residues 289-294, and 4.9 Å for residues 325-516. The median structures of the best ten scoring models are shown in Fig. 2A. Using the low-pass filtered map, we could also manually trace the main chain for nt −9 to −3 and nt +9 to +13 of pre-mRNA. Combined with the previous atomic model, 23 nucleotides of the pre-mRNA were manually built, of which the upstream could directly

insert into NCBP1/NCBP2 heterodimer and the downstream could interact with the RRM2 domain of Nam8.

The model of the pre-RNA helix and the putative localization of the BBP/Mud2 binding region were based on the 3.2Å resolution structure of the Act1 complex. The modeling procedure is similar to that used for modeling the Ubc4 complex except for the following differences. Firstly, we could observe the density for ~25bp double RNA helix with clear major and minor grooves on a binding surface formed by the C-terminal tail of Prp39, the N-terminal domain of Prp42, and the C-terminal domain of U1C. Such double helix density was also observed in the pre-B complex structure and was tentatively modeled as part of U2 snRNA. Since our Act1 complex was assembled from *in vitro* transcribed Act1 pre-mRNA, purified U1 snRNP, BBP, and Mud2 proteins, there is no U2 snRNA present in our sample (Extended Data Fig. 5B). Although some bases can be separated in the density map of this double helix, we were unable to model specific nucleotides. Nonetheless, there is weak density connecting it to the 5' ss, suggesting that this double helix belongs to the pre-mRNA region downstream of the 5' ss. Secondly, there is a large volume of weak density close to pre-mRNA double helix (Extended Data Fig. 4I). The density can be best interpreted as the BBP/Mud2 dimer bound to pre-mRNA, given that its location corresponds roughly to where U2 SnRNP is in the A complex structure (Extended Data Fig. 4I).

The above models were refined using PHENIX in real space [51] with secondary structure and geometry restraints. Refinement statistics of the E complex were summarized in Extended Data Table 1. These models were also evaluated based on Morprobity scores [52] and Ramachandran plots (Extended Data Table 1). Model/map FSC validation was shown in Extended Data Fig. 2g **and** 3g. Representative densities for the proteins and RNA are shown in Extended Data Fig. 4. All structure-related images in this paper were generated using UCSF CHIMERA [46] and CHIMERAX [53].

To determine the minimum number of nucleotides needed to connect an upstream BPS to a downstream 5' SS in the A complex[14], we modeled connection lengths ranging from 21 to 30 nucleotides between nucleotides 74 and −1 of pre-mRNA (chain I in PDB ID 6g90, nucleotide 70 is BP and +1 is 5' ss) using the Rosetta RNP-denovo method with full-atom refinement [28,50]. Nucleotides 75-78 of pre-mRNA were excised from the structure and all other nucleotides were kept fixed. 21-30 uridines were modeled de novo to connect nucleotide 74 to nucleotide −1. During the initial low-resolution stages of the modeling, score terms rewarding favorable RNA-protein interactions, RNA base pairing, and compact RNA structures were turned off. Score terms that penalize clashes within the RNA and between the RNA and protein were included during this stage. During the final all-atom refinement, the complete all-atom RNA-protein score function was used. The weight on the score term penalizing chainbreaks was increased to 50.0 and models with a chainbreak score less than 0.5 were considered fully connected. The top scoring model (at least 250 models were built for each connection length) by full Rosetta score was used as a representative model for each connection length. We found that the representative model for 22-nucleotide connection length is fully connected and has similar total Rosetta score as models for longer connection lengths, indicating 22 nucleotides are sufficient for connecting nucleotide 74 to −1 without highly unfavorable interactions such as clashes.

Crosslinking and mass spectrometry

Purified yeast spliceosome E complex was crosslinked with 10 mM DSSO (disuccinimidyl sulfoxide) for 45 min at 4 °C, the reaction was quenched by adding ammonium bicarbonate to a final concentration of 50 mM. Crosslinked complex was proteolytically digested according to the FASP (filter-aided sample preparation) protocol as previously described [54]. Briefly, ~100 μg of crosslinked sample was reduced, alkylated, and digested at 1:50 with sequencing grade trypsin (Promega) by incubating at 37 °C for 18 hours. Peptides were eluted and acidified to 0.1 % formic acid. Enrichment of crosslinked peptides was performed by using strong cation exchange chromatography (SCX) with a Dionex UltiMate 3000 system (Thermo Fisher Scientific). A Proteomix SCX-NP1.7 column (4.6 mm inner diameter, 150 mm length, Sepax Technologies) was used. Briefly, peptides were separated using the following gradient: 0 % B (0 – 3.5 min), 0 – 22.5 % B (3.5 – 18.5 min), 22.5 – 50 % B (18.5 – 21.5 min), 50 – 100 % B (21.5 – 23 min), 100 % B (23 – 25.5 min) with solvent A (10 mM KH_2PO_4, 25 % acetonitrile, pH 3.00) and solvent B (10 mM KH_2PO_4, 25 % acetonitrile, 500 mM KCl, pH 3.00) at a flow rate of 0.7 ml/min. Fractions were collected every minute. Fractions 6-26 were pooled into groups of three and desalted using StageTips for subsequent LC-MS/MS analysis.

Crosslinked peptides were then analyzed by nano-UHPLC-MS/MS (Easy-nLC1200, Orbitrap Fusion™ Lumos™Tribrid™, Thermo Fisher Scientific). 14 μl of sample was directly loaded onto an in-house packed 100 μm i.d. × 250 mm fused silica column packed with CORTECS C18 resin (2.7 μm, spherical solid-core). Samples were run at 400 nL/min over a 90 min linear gradient from 4-32% acetonitrile with 0.1% formic acid. The mass spectrometer was operated in positive ion mode with two sequential experiments per duty cycle. For crosslink peptide identification, MS1 scans were ran in the orbitrap from 375-1500 m/z at 60,000 resolution. MS2 was performed on the top 4 ions from each precursor scan and fragmented at a CID collision energy of 22%. MS3 was triggered by the targeted mass difference of 31.9721 Da represented by the cleavage of the DSSO sulfoxide bond, and was performed as a stepped HCD collision energy of 33% +/−3. For linear peptide identification, a second precursor scan was performed at 120000 resolution in a scan range of 350-1000 m/z. Stoichiometric sampling of ions for MS2 fragmentation was capped at 2 seconds and performed at an HCD collision energy of 30% in the orbitrap. Data acquisition was performed using Xcalibur (version 4.1) software.

Instrument raw files were directly loaded in to Proteome Discoverer 2.2 and were searched against twenty-two proteins making up the E Complex of U1snRNP from *S. cerevisiae* of the Swiss-prot database (update 2018_08_08) using the XlinkX plugin. Search parameters included carbamidomethylation-C as a fixed modification, oxidation-M, DSSO-K, DSSO/amidated-K, and DSSO/hydrolysed-K as variable modifications, allowing for two missed cleavages. MS2_MS3 was set for crosslink detection against DSSO. Precursor mass tolerance was set to 10 ppm, with MS/MS mass tolerance set to 20 ppm. Results were manually validated and visualized using xVis [55].

Oligo-directed RNase H digestion of pre-mRNA in purified E complex

Purified E complex with Act1 or Dyn2 IEI pre-mRNA as substrate was incubated with RNase H (New England Biolabs) in the presence or absence of 5 μM DNA oligo, at 25 °C for 30 min. The complex was then bound to amylose resin pre-washed with buffer G120. The resin was washed with buffer G120 and eluted in buffer G120 containing 10 mM maltose. The eluted samples were analyzed on SDS-PAGE and stained with Coomassie to visualize the proteins. For RNA detection, the samples were digested with 1 μg/μL proteinase K and separated on 7M urea denaturing polyacrylamide gel and stained with EtBr, or on native agarose gel and stained with SYBR gold (Life Technologies). DNA oligo used for digestion of Act1 is 5'-AAAATAAACGATGACACAG-3', and for Dyn2 is 5'-TCATGGAAGAAAACCTCAC-3'.

Chase experiment with assembled E complex in U1-depleted yeast extract

BSY82 (GAL-U1) yeast strain [56] obtained from Dr. Michael Rosbash's lab was maintained in YEP media containing 2% galactose. For U1 snRNA depletion cultures, log phase cells growing in 2% galactose were diluted into 2% glucose containing media to an OD600 of 0.03 and grown for 17 h to an OD600 of 2.5. Yeast splicing extracts were prepared from 2 liters of yeast cells cultured in media containing either galactose or glucose. Splicing reactions were carried out at 23°C for 15 min in a 25 μL reaction containing 2.5 nM M3-Act1 pre-mRNA alone or purified E complex, with or without 50-fold (in molar quantity) of Act1-M3 (Act1 pre-mRNA containing 54nt 3' exon fused with three copies of MS2 stem loops at the 3′-end), 40% yeast extract, and splicing buffer (60 mM potassium phosphate, pH 7.4, 3% PEG 8000, 2.5 mM $MgCl_2$, 2 mM ATP). RNAs were then phenol/chloroform extracted and precipitated with 2.5 volumes of ethanol. After DNase I (Roche) treatment, first strand cDNAs were synthesized from 1 μg of RNA using ProtoScript II reverse transcriptase with reverse primers specific to M3-Act1, U1 snRNA, or U2 snRNA. PCR was performed using cDNA transcribed from 25 ng of RNA as template and the following primers: MS2 Forward 5'-TCCGATATCCGTACACCATC-3'; Act1 exon 2 Reverse 5'-TGATACCTTGGTGTCTTGGTCT-3'; yeast U1 Forward 5'-AAACATGCGCTTCCAATAGT-3', Reverse 5'-TATGTGTGTGTGACCAAGGAG-3'[57](*75*); and yeast U2 Forward 5'-AACTGAAATGACCTCAATGAGGCTC-3', Reverse 5'-AGACCTGACATTAGCGGAAAACAAC-3'. The products were analyzed on 3% low melting point (LMP) agarose gel stained with EtBr.

The same experiment was also performed using EFM5 IEI-101-M3 or HMRA1 IEI-246-M3 pre-mRNA and E complex assembled on both pre-mRNAs. A second EFM5 IEI-101 pre-mRNA was designed to remove the primer binding sites so it is invisible in the RT-PCR reaction and was *in vitro* transcribed to be used as competing RNA. For HMRA1, wild type IEI-246 pre-mRNA was used as competing RNA. Primers used to detect circRNA formed specifically from EFM5 IEI-101 are: IEI-101 cir Forward 5'-CCTTGAAGAATTCAAAAGAGAGGATAG −3' and cir Reverse 5'-CGAGGGCATTAGCAGAAAG −3'. Primers used to detect circRNA formed specifically from HMRA1 IEI-246 are: IEI-62 cir Forward 5'-TCCACTTCAAGTAAGAGTTTGG-3' and cir Reverse 5'- GAGTAAAGTAGTGGCATTAGTCA-3'.

Analyzing the role of Act1 intronic secondary structure in splicing

We mutated multiple stretches of sequences in the 5' ss to BPS region of Act1 so that it no longer forms extensive secondary structures. The final sequence of this region (mutated nucleotides underlined) is: GTATGTTCTAGCGCTTGCACCATCCCATTTAACTGTAAAAAAAAATGCACGGTCCC AATTGCTCGAGAGATTTCTCTTTTACAAAAAAATACTATTAAAAAAAGAGAAAAA ACCTCCTATATTGACTGATCTGTAATAACCACGATATTATTGGAATAAATAGGGGCT AAAAATTTGGAAAAAAAGAAAAACTGAAATATTTTCGTGATAAGTGATAGTGATA AAAAAAAATTATTTGCTACTGTGTCTCATGTACTAA. We synthesized this mutant region through Genscript and replaced the 5' ss to BPS region of the Act1-Cup1 reporter plasmid [19] with this sequence to generate the mutant reporter plasmid pMA6.

The WT and mutant Act1-Cup1 reporter plasmids were transformed into yeast strain BY4741, and grown in synthetic complete (SC) -Leu medium until OD600=1.0. RNA is extracted from 5 mls of culture, treated with DNase I (Roche), and reverse transcribed using the ProtoScript II First Strand cDNA Synthesis Kit (New England Biolabs) and random primers. qPCR was performed using the iTaq™ Universal SYBR® Green Supermix (Biorad) with cDNA transcribed from 10 ng of RNA as template. The primers for detecting pre-mRNAs are located in the Act1 intron and Cup1: Act intron Forward 5'-TTATTTGCTACTGTGTCTCATG-3'; YAC7 Reverse 5'-GCATTGGCACTCATGACCTT-3'. The primers for detecting total reporter mRNA are located in Act1 exon 2 and Cup1: ActEx2 Forward 5'- GTTCTGGTATGTGTAAAGCC-3'; CUP1end Reverse 5'-CCAGAGCAGCATGATTTCTT-3'.

Co-purification assay in yeast

The coding regions of yeast Prp40, Snu71, and Luc7 full length or truncations were amplified by PCR using genomic *S. cerevisiae* DNA as templates, and ligated into pRS414, pRS416 and pRS317 vectors. The final plasmids constructed are: pRS414/GPD-protA-Prp40 (full-length, FF1-3 and FF4-6), pRS317/GPD-Prp40, pRS414/GPD-protA-Snu71, pRS416/GPD-CBP-Luc7 or Luc7ΔCC (Luc7 coiled coil domain (residues 123-190) deletion). Yeast BCY123 cells were transformed with different combination of the plasmids and selected on appropriate selective media. Clones from the transformation were cultured in 50 ml of liquid selective media to OD600=3-4. Cells were harvested and lysed in lysis buffer (50 mM Tris-HCl, pH 8.0, 150 mM NaCl, 0.05% NP40, 1 mM DTT, 2.5mM $CaCl_2$, 1.5mM $MgCl_2$) using the bead-beating method. The lysates were incubated with IgG resin for 3 hr at 4°C. The resins were washed with the lysis buffer. The proteins were cleaved off IgG resin using TEV protease in buffer (20 mM Tris-HCl, pH 8.0, 150 mM NaCl, 0.01% NP40, 0.5mM DTT). The proteins were separated on SDS-PAGE and transferred to a nitrocellulose membrane. Western blot was performed using an anti-CBP tag antibody (GenScript, A00635).

Dyn2 splicing analyses

The *DYN2* gene was PCR amplified from *S. cerevisiae* genomic DNA along with the 5′-UTR (356 bp) and 3′-UTR (305 bp) and cloned into the pRS415 vector. The BPS of the first intron was mutated from "TACTAAC" to "TAGTACC" and the 5'SS of the second intron

from “GT” to “CG” separately or together, to generate the I1-BP, I2-5’SS, and double mutant. The wild type and mutant plasmids were transformed into a *DYN2* deletion yeast strain (Open Biosystem), and transformants were selected on SC -Leu plates. Cells were grown in SC -Leu media to OD600 of ~1.0. Total RNA was isolated from 10 ml of cells using a hot-phenol extraction method and dissolved in 100 μl of diethylpyrocarbonate (DEPC)-treated water. A total of 1 μg of RNA was treated with DNase I and reverse transcribed into cDNA. qPCRs were performed using the following primers: E1 Forward 5’-CCAAAATGAGCGATGAAAATAAGAG-3’ and I1 Reverse 5’-TCATGGAAGAAAACCTCACTC-3’ to detect exon 1-intron 1 product; I1 Forward 5’-TATGTCAGTTAATCTCAGTCACAAT-3’ and E2 Reverse 5’-TATGTCAGACGCCTTAACAATAG-3’ to detect intron 1-exon 2 product; E2 Forward 5’-CTATTGTTAAGGCGTCTGACATA-3’ and I2 Reverse 5’-GGTCTAAGTTTTCTCCTTGTTAG −3’ to detect exon 2-intron 2 product; I2 Forward 5’-CATGTTTTGTGTGTGTACATTTG-3’ and E3 Reverse 5’-CAGGTATTGCCGTATTTGAC-3’ to detect intron 2-exon 3 product; E3 Forward 5’-CGACAAGCTGAAAGAGGATA-3’ and E3 reverse to detect exon 3.

CircRNA detection from purified spliceosome

Spliceosome was purified from three liters of yeast cells carrying the $Prp22^{H060A}$ mutant to enrich the post-catalytical complex [20] to increase our chance of detecting branching and ligation products before their release. RNA from the purified spliceosomal complex was purified and reverse transcribed into cDNA. PCR was performed to detect the presence of T-branches and circRNAs from yeast multi-intronic genes *EFM5* (YGR001C) and *HMRA1* (YCR097W), using circle-specific primers [7] and the following primers for T-branches: EFM5 I1 Forward 5’- TTTTCAACACAGTAACGTAGAATTAC-3’, I2 Reverse 5’-AACAGTTAGTAAGATGAAAAGATACTGG-3’; HMRA1 I1 Forward 5’-GTATGTTTTCATTTCAAGGATAG-3’, I2 Reverse 5’-TGTTAGTATAGGATATATTTAAGTTTGA-3’. PCR products were analyzed on 3.5% LMP agarose gel stained with EtBr, and cloned into pMiniT vectors (New England Biolabs) for sequencing.

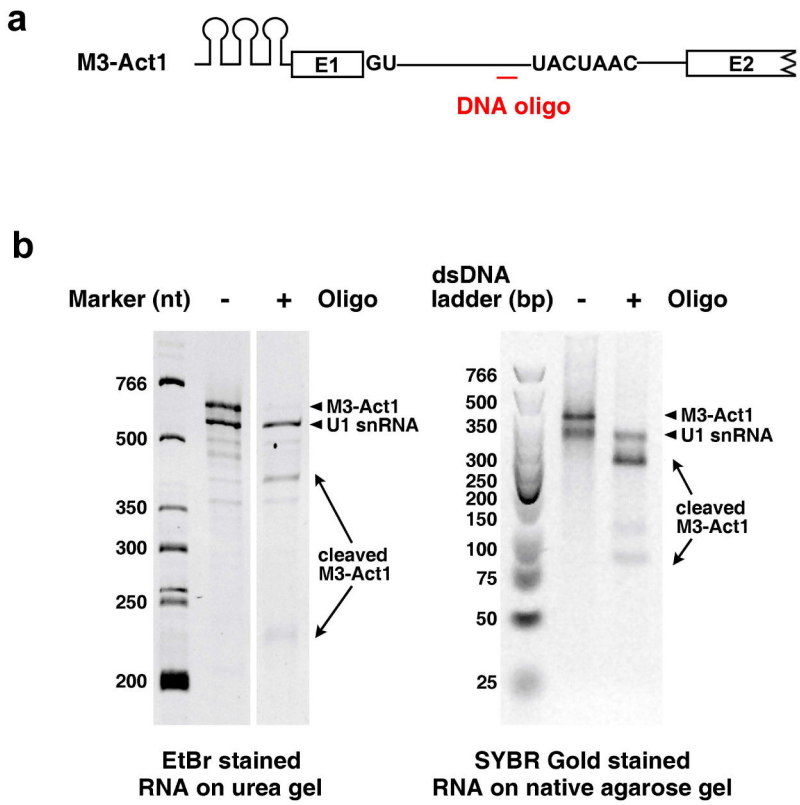
CircRNA detection from EFM5 and HMRA1 IEI constructs in vivo

The region from +47 to +630 of *EFM5* gene (containing EFM5 partial intron 1, exon 2, and partial intron 2) was PCR amplified from *S. cerevisiae* genomic DNA and was inserted between a GPD promoter and a CYC1 terminator, and the resulting expression cassette was cloned into a pRS317 vector, generating the EFM5 IEI plasmid. The BPS of the first intron was mutated from “TACTAACTAAC” to “TAGTACCTACC” (mutations underlined) [58] and the 5’SS of the second intron from “GT” to “CG” separately, to generate the BP and 5’SS mutant. The IEI-101 truncation was generated by fusing the first 49 nt to the last 52 nt of the middle exon. The IEI-63 truncation was engineered on IEI-101 to shorten the middle exon to 63 nt in length, of which the sequence is: 5’-GACACTTTCTGCCCTTGAAGAATTCAAAAGAGAGGATAGATTGTTAATTGACCCAAACAAAGT-3’.

The EFM5 IEI plasmid and the empty pRS317 vector were transformed into an EFM5 deletion yeast strain (Open Biosystem). Cells were grown in SC -Lys media to OD 600 of ~1.0. Total RNA was isolated and reverse transcribed into cDNA. For RNase R treatment, 1 μg of total RNA was incubated at 37 °C for 30 min with 5 U/ μg of RNase R (Epicentre Technologies) and used directly for reverse transcription without further purification. PCR was performed using specific primers [7] to detect circRNA formed from exon2 of EFM5 using the following primers: EFM5 cir Forward 5'-GAGAGGATAGATTGTTAATTGACCC-3' and EFM5 cir Reverse 5'-CTTTTGAATTCTTCAAGGGCA-3'. The primer pair used to detect un-spliced EFM5 pre-mRNA are: EFM5 I1 Forward 5'-TTTTCAACACAGTAACGT AGAATTAC-3', I2 Reverse 5'- GAGTAGGGATATGTTTATGATATACATAC −3'. The products were analyzed on 3% LMP agarose gel stained with EtBr.

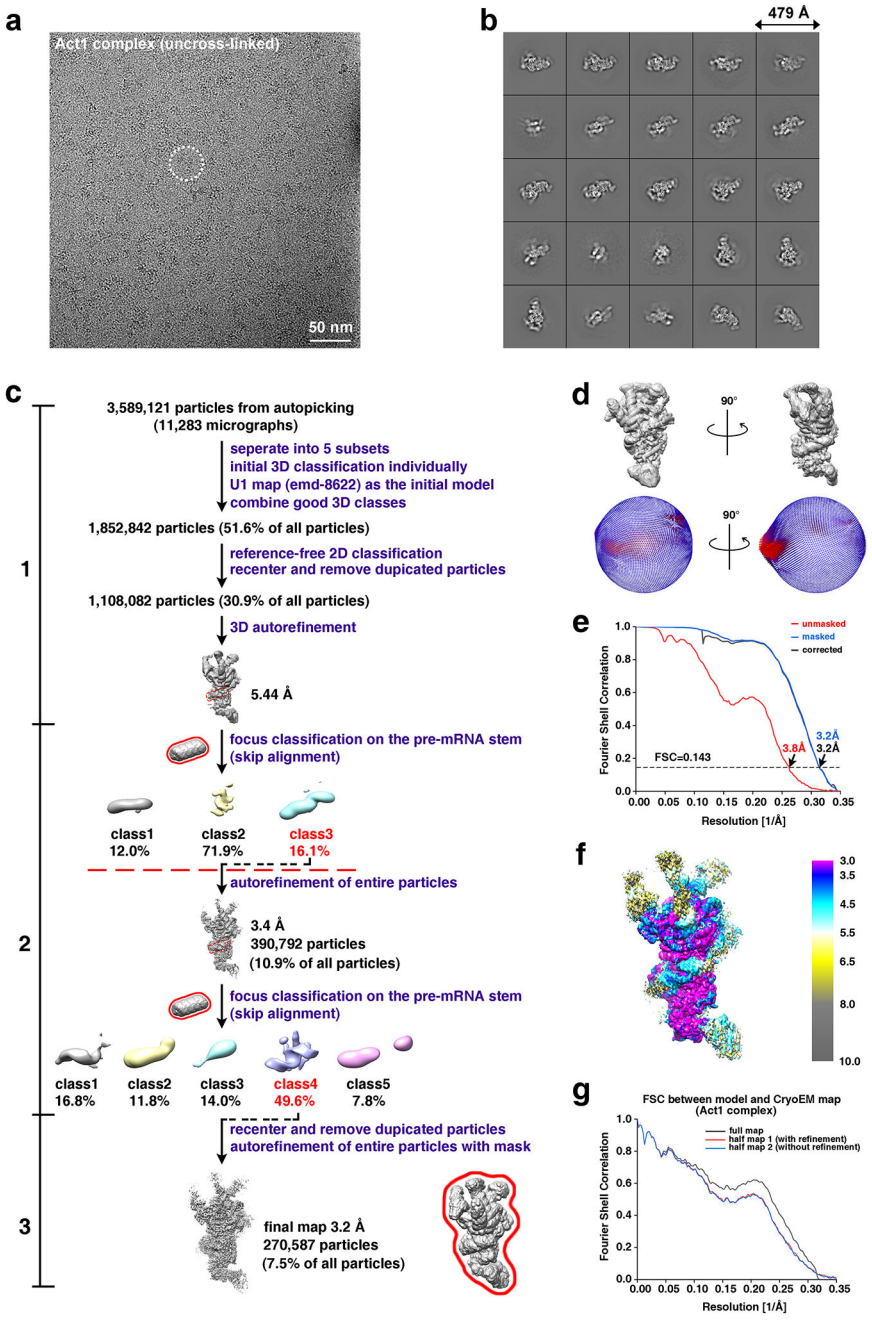
The region from +116 to +439 of *HMRA1* gene (containing exon 2 flanked by partial intron 1 and intron 2) was PCR amplified from *S. cerevisiae* genomic DNA and cloned into pRS317 vector in the same way. The IEI-62 truncation was engineered to shorten the middle exon to 62 nt in length, of which the sequence is: 5'-TTTATAATGGAAAGTAATTTGAC TAATGCCACTACTTTACTCCACTTCAAGTAAGAGTTTGG −3'. Primers used to detect circRNA are IEI-62 cir Forward and Reverse primer, which are the same as those used for IEI-246. The primer pair used to detect un-spliced HMRA1 pre-mRNA are: HMRA1 I1 Forward 5'-CCA AGAACTTAGTTTCGACTCTAGATTTCAAGGATAGCCTTTGAATC-3', I2 Reverse 5'-AACTAATTACATGATGGGCCCGGATATATTTAAGTTTGATTCTCATATTACATAC-3'.

Extended Data



Extended Data Figure 1. *In vitro* assembly and purification of the Act1 complex.

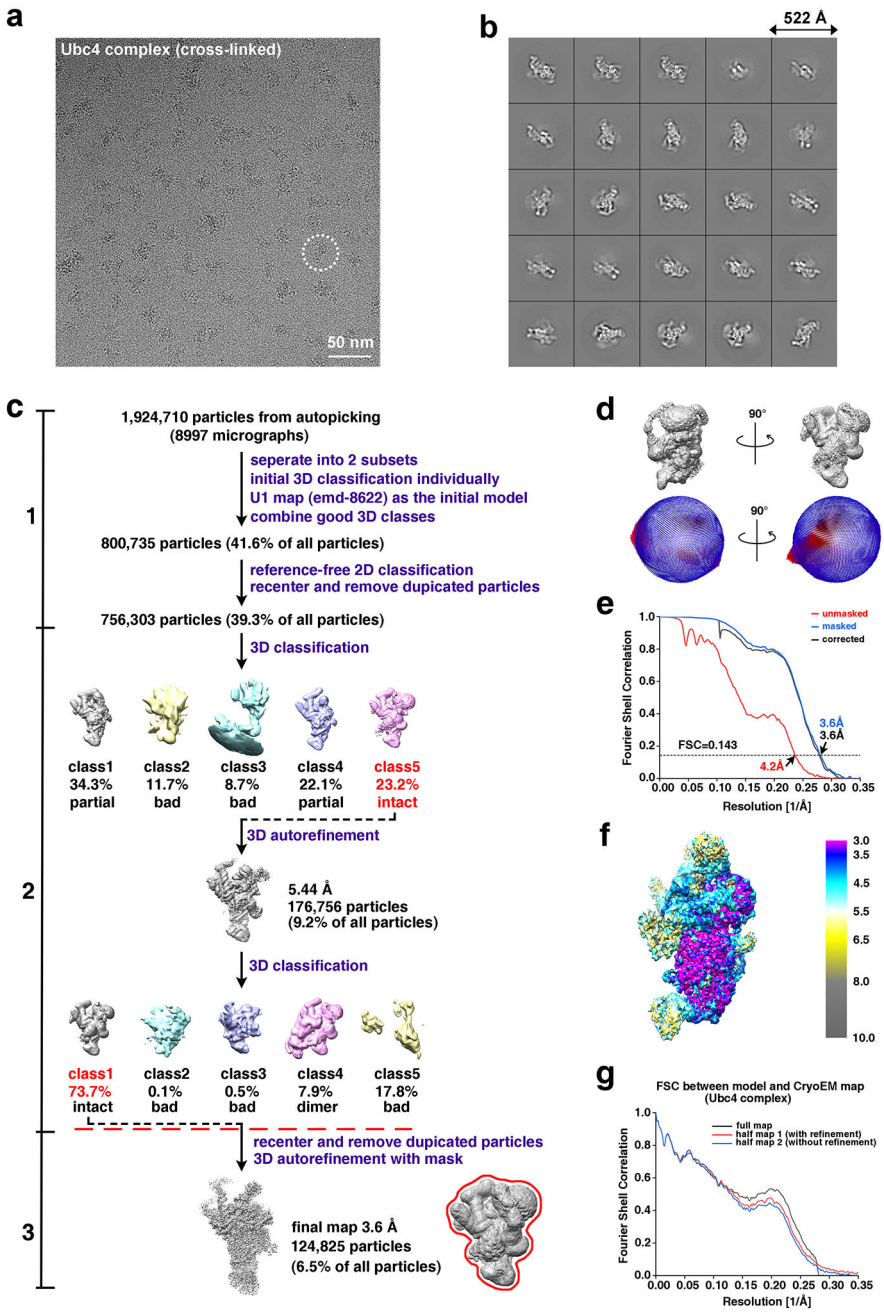
(a) A schematic representation of the Act1 pre-mRNA tagged with three MS2-binding sites (M3-Act1) used for E complex assembly and purification. Boxes represent exon 1 (E1) and truncated exon 2 (E2). The 5' ss (GU) and BPS (UACUAAC) are also shown. The red line represents the DNA oligo complementary to a region 5nt upstream of the BPS for the RNase H cleavage experiment. **(b)** RNA components of the assembled E complex (with or without DNA oligo and RNase H treatment) after proteinase K digestion are shown on a denaturing urea gel or native agarose gel. These results demonstrate that RNase treatment cleaved M3-Act1 into two fragments. Note that the sizes of RNA on the native gel do not match their linear length, possibly due to the existence of secondary structures. This experiment was repeated two additional times with similar results.



Extended Data Figure 2. The CryoEM structural determination process for the Act1 complex.

(a) A representative drift-corrected cryoEM micrograph (out of a total of 11,283 images) of the E complex assembled on the Act1 pre-mRNA. A representative particle is shown in a white dotted circle. **(b)** Representative 2D class averages of the Act1 complex obtained in RELION. This experiment was repeated one additional time with similar results. **(c)** Data processing workflow. For processing above the red dash line, the particle images were binned to a pixel size of 2.72 Å. The rest of processing was performed with a pixel size of 1.36 Å. The masks used in data processing are outlined with red solid line. Please refer to Methods for more details. **(d)** Angular distribution for all particles used for the final 3.2 Å

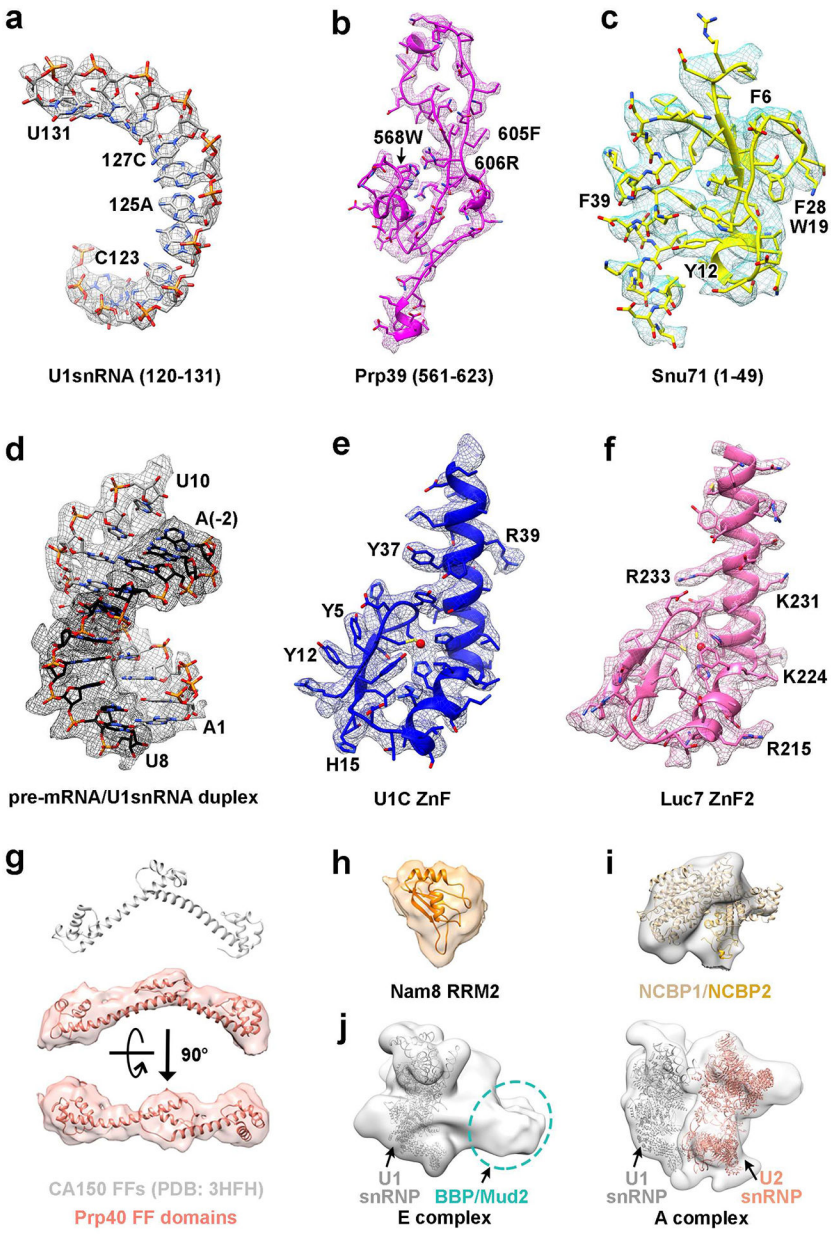
map of the Act1 complex. **(e)** FSC as a function of spatial frequency demonstrating the resolution for the final reconstruction of the Act1 complex. **(f)** Resmap local resolution estimation. **(g)** FSC coefficients as a functional of spatial frequency between model and cryoEM density maps. The generally similar appearances between the FSC curves obtained with half maps with (red) and without (blue) model refinement indicate that the refinement of the atomic coordinates did not suffer from severe over-fitting.



Extended Data Figure 3. The CryoEM structural determination process for the Ubc4 complex.

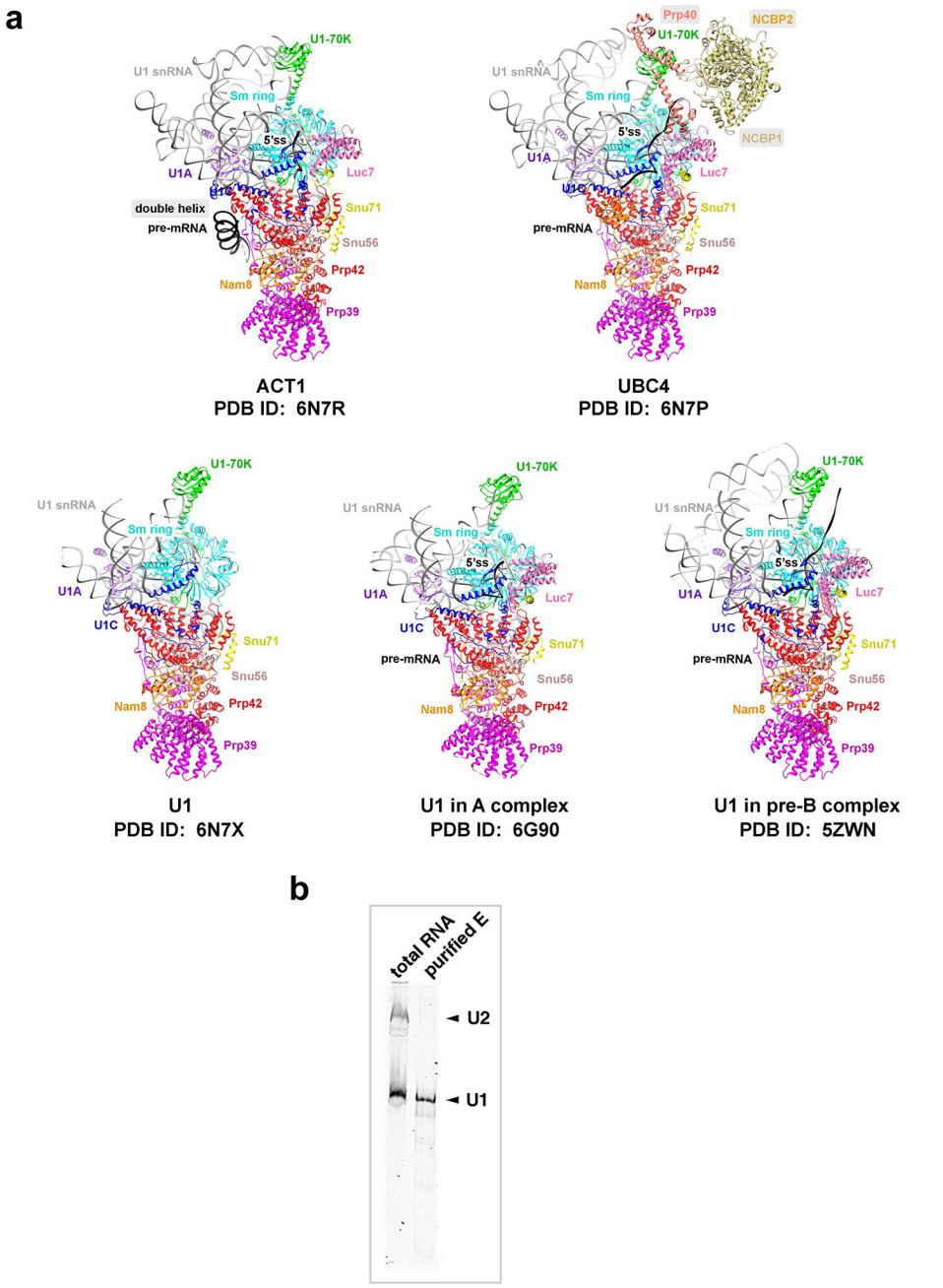
(a) A representative drift-corrected cryoEM micrograph (out of a total of 8,997 micrographs) of the E complex assembled on the Ubc4 pre-mRNA. A representative particle is shown in a white dotted circle. **(b)** Representative 2D class averages of the Ubc4 complex obtained in RELION. **(e)** 2D classification of negative-stain TEM images of the E complex assembled on Dyn2 IEI pre-mRNA. This experiment was repeated one additional time with similar results. **(c)** Data processing workflow. For processing above the red dash line, the particle images were binned to a pixel size of 2.72 Å. The rest of processing was performed with a pixel size of 1.36 Å. The masks used in data processing are outlined with red solid line.

Please refer to Methods for more details. **(d)** Angular distribution for all particles used for the final 3.6 Å map of the Ubc4 complex. **(e)** FSC as a function of spatial frequency demonstrating the resolution for the final reconstruction of the Ubc4 complex. **(f)** Resmap local resolution estimation. **(g)** FSC coefficients as a functional of spatial frequency between model and cryoEM density maps. The generally similar appearances between the FSC curves obtained with half maps with (red) and without (blue) model refinement indicate that the refinement of the atomic coordinates did not suffer from severe over-fitting.



Extended Data Figure 4. Representative cryoEM density maps of the E complex.

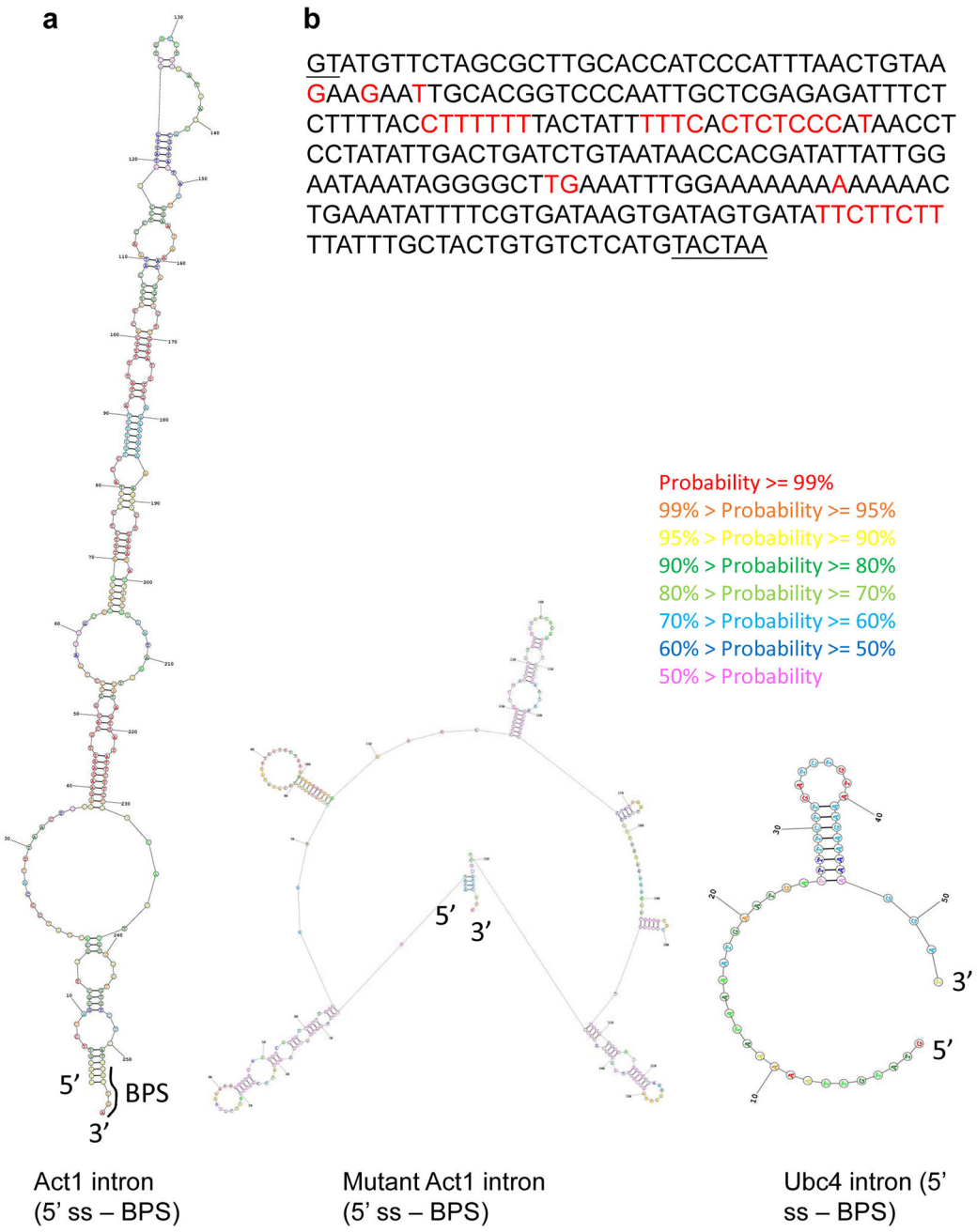
Panels (a-i) are densities for the Ubc4 complex and (j) is density for the Act1 complex. The cryoEM density maps are shown for **(a)** selected regions of U1 snRNA; **(b)** C-terminal region of Prp39; **(c)** N-terminal domain of Snu71; **(d)** the pre-mRNA and U1 snRNA duplex; **(e)** U1C ZnF domain; **(f)** Luc7 ZnF2 domain; **(g)** the tandem FF domains of Prp40 (known structures of tandem FF domains from CA150 are also shown with the characteristic boomerang-shape); **(h)** the RRM2 domain of Nam8; **(i)** NCBP1 and NCBP2; **(j)** the weak density in the Act1 complex that is assigned as the putative BBP/Mud2 heterodimer. The A complex is also shown, with U1 snRNP in the same orientation as the Act1 complex and U2 snRNP located in similar positions as the BBP/Mud2 heterodimer with respect to U1 snRNP. The map of Act1 complex was low-pass filtered to 40 Å.



Extended Data Figure 5. Structural and biochemical characterization of the Act1 and Ubc4 complexes.

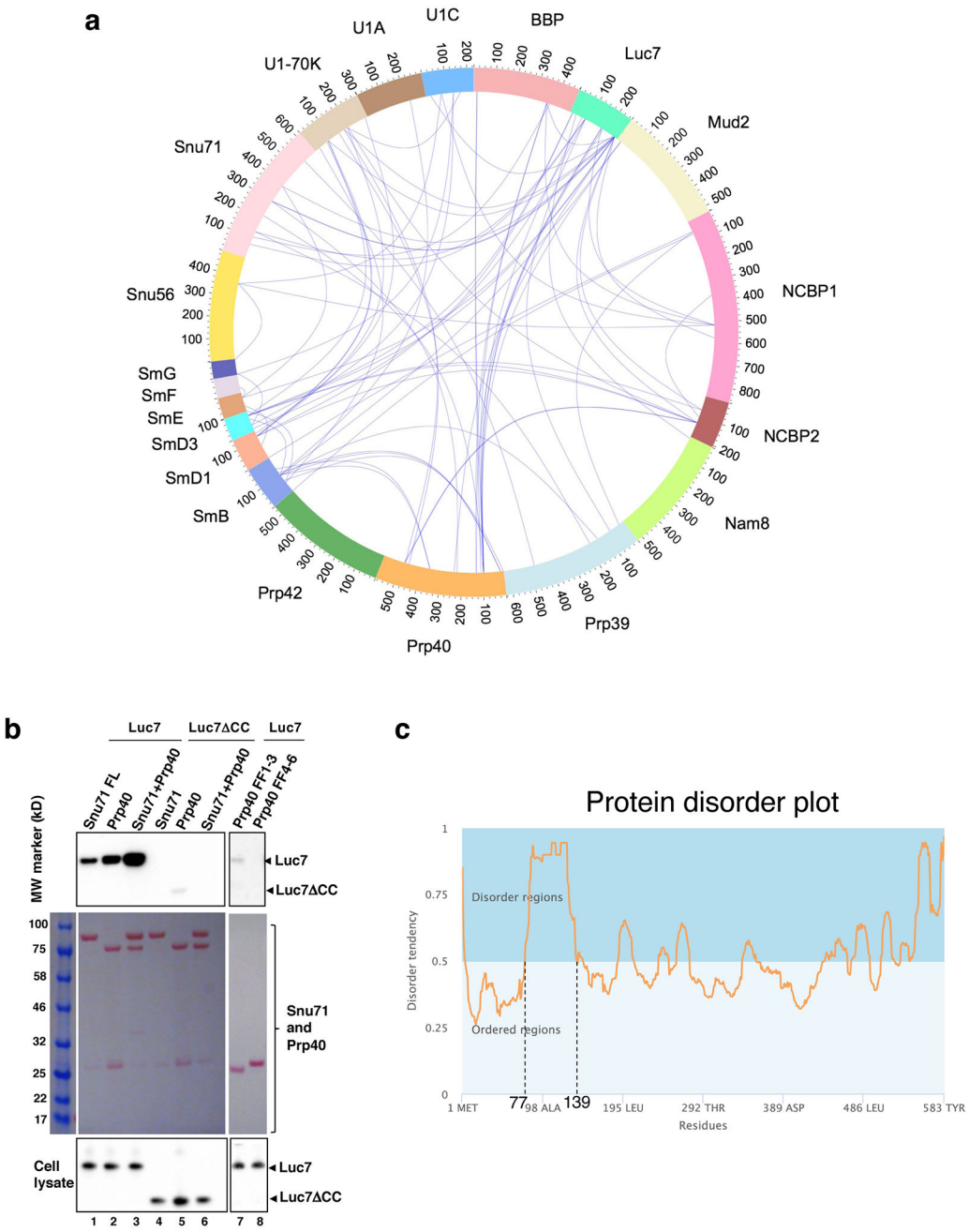
(a) Comparison of the ribbon models of the Act1 complex, the Ubc4 complexes, and U1 snRNP from other previously determined structures (the U1 snRNP, A, and pre-B complex). Labels in shade indicate protein or RNA components that are different between the Act1 and Ubc4 complexes. These components and the RRM2 domain of Nam8 are also absent from previously determined structures. Note that U1-70K is shifted towards NCBP2 in the Ubc4 complex. **(b)** Purified E complex does not contain U2 snRNA. A native polyacrylamide gel shows the solution hybridization (*78*) result of total cellular RNA or RNA from purified E

complex hybridized with fluorescent probes specific for U1 and U2 snRNAs. This experiment was repeated one additional time with similar results.



Extended Data Figure 6. Secondary structures in the region between the 5' ss and BPS in the WT and mutant Act1 and Ubc4 pre-mRNAs.

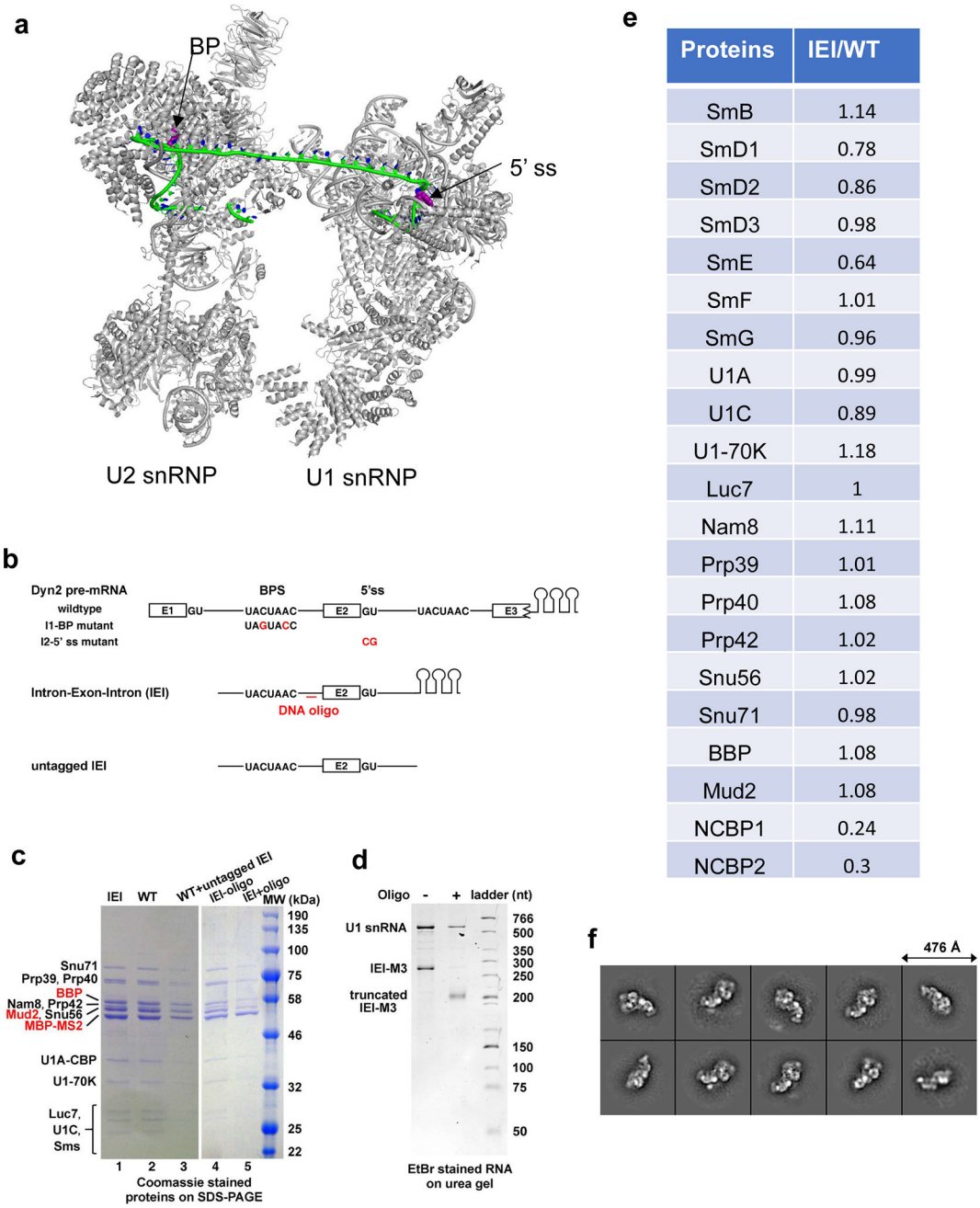
(a) Secondary structures predicted by RNAstructure 6.0 (https://rna.urmc.rochester.edu/RNAstructureWeb/). **(b)** Sequence between the 5' ss and BPS (underlined) of Act1. Red nucleotides were mutated to A (other than the one A which was mutated to G) in the mutant Act1 to disrupt predicted secondary structures.



Extended Data Figure 7. Protein interactions in the Ubc4 complex.

(a) DSSO crosslinking and mass spectrometry analyses of the Ubc4 complex. Each blue line indicate crosslinks observed between a pair of Lys residues. Note that BBP/Mud2 are crosslinked to Luc7, Prp40, Snu56, and Snu71. **(b)** Co-purification assays probing the interaction between Snu71 (or Prp40) and Luc7. Various combinations of protein A-TEV-Prp40, protein A-TEV-Snu71, and CBP-tagged Luc7 or Luc7ΔCC [coiled coil domain (residues 123-190) of Luc7 deleted] were co-overexpressed in yeast (only Snu71 is protein A tagged in the Snu71+Prp40 lanes), purified using IgG resin, eluted through TEV cleavage, analyzed on SDS-PAGE, and visualized using both Western blot with an anti-CBP antibody

to detect Luc7 (top) and Ponceau S stain to show Snu71 or Prp40 (middle). Western blot using the same anti-CBP antibody was used to demonstrate Luc7 expression levels in cell lysates (bottom). The faint band around 26 kD in all lanes is TEV. This experiment was repeated one additional time with similar results. **(c)** The linker (residues 73-131) between the WW and FF domains of Prp40 is predicted to be disordered using program *MetaDisorderMD2* (*79*).



Proteins	IEI/WT
SmB	1.14
SmD1	0.78
SmD2	0.86
SmD3	0.98
SmE	0.64
SmF	1.01
SmG	0.96
U1A	0.99
U1C	0.89
U1-70K	1.18
Luc7	1
Nam8	1.11
Prp39	1.01
Prp40	1.08
Prp42	1.02
Snu56	1.02
Snu71	0.98
BBP	1.08
Mud2	1.08
NCBP1	0.24
NCBP2	0.3

Extended Data Figure 8. Computational and biochemical characterization of the EDC.

(a) The minimal length of RNA needed to connect the upstream BP and downstream 5' ss in the A complex is modeled using the Rosetta RNP-denovo method. The A complex (PDB ID 6g90) is shown in grey. The pre-mRNA is shown in green. The upstream BP and downstream 5' ss are shown in purple space filling models. 28 nucleotides are sufficient to connect the upstream BP and downstream 5' ss (not including the BP and 5' ss themselves) without any chainbreak and clashes. **(b)** Schematics of the Dyn2 pre-mRNA WT and mutants (mutated nucleotides shown in red), IEI, and untagged IEI used for the EDC assembly and *in vivo* exon definition experiments. Stem-loops represent the MS2 binding

sites, and the red line represents the DNA oligo used for RNase H cleavage. **(c)** SDS-PAGE shows protein components of complexes assembled on WT and IEI substrates (lanes 1-2), on WT in the presence of competing untagged IEI (lane 3), and on IEI after RNase H treatment in the absence and presence of the DNA oligo (lanes 4-5). This experiment was repeated one additional time with similar results. **(d)** RNA components of the same complexes as in lanes 4-5 of (b), confirming that RNase H treatment + oligo indeed cleaves the pre-mRNA. The smaller cleaved fragment (61 nucleotides) is difficult to see since EtBr has a low efficiency staining short single stranded RNA. This experiment was repeated two additional times with similar results. **(e)** Mass spectrometry analyses of spliceosome assembled on the Dyn2 IEI and WT pre-mRNA indicate that the two complexes have the same components in similar quantities with the exception of NCBP1 and 2 which are absent from the IEI complex. **(f)** 2D classification of negative-stain TEM images of the E complex assembled on Dyn2 IEI pre-mRNA. This experiment was repeated one additional time with similar results.

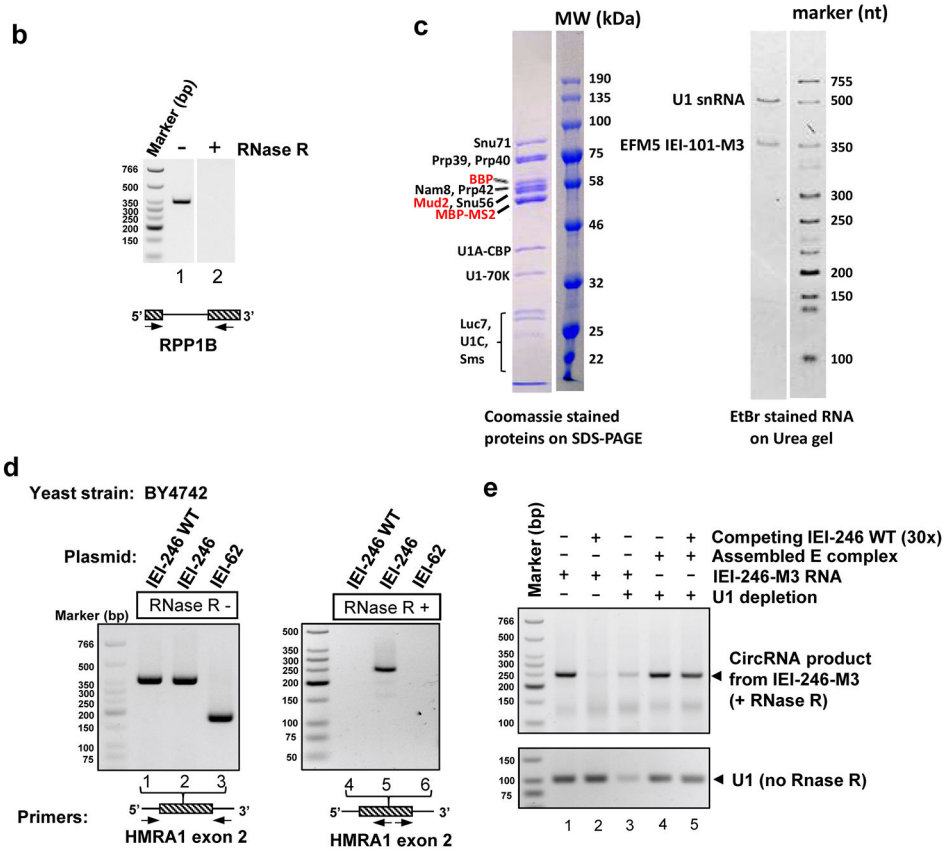
a

EFM5 PCR product generated from circRNA using outward facing primers on exon2 (Fig. 6A, lane 9):
…tatcagAACCACTGGACTTCAGTGATGAAATTAAAGGAAAAGTTGATAGATTGTTAATTGACCCACCTTTTT
TAAATGAAGATTGTCAAACAAAGT/GACACTTTCTGCTAATGCCCTCGCTGCCCTTGAAGAATTCAAAAGAGA
GGAACAACAACATCAAGAAGCCTTTCAAAAGCTTTACGACGatcatg…

EFM5 PCR product generated from T-branch (Fig. 6A, lane 1):
…catgatTTTTCAACACAGTAACGTAGAAT**TACTAACT--lGT**ATGTATATCATAAACAT
ATCCCTACTCATTTTTTAATCTTTTTTCCAGTATCTTTTCATCTTACTAACTGTTctgata…

HMRA1 PCR product generated from circRNA using outward facing primers on exon2 (Fig. 6A, lane 10):
…tatcagGAAAGCAAAGCCTTAATTCCAAGGAAAAAGAAGAAGTTGCAAAGAAATGTGGCATTACTCCACTTC
AAGTAAGAGTTTGG/TTTATAATGGAAAGTAATTTGACTAAAGTAGAGCAACATACATTACACAAAAATATTT
CTAACAATAGGTTAGAAATATACCACCACATTAAAAAAGAGAAGAGCCCAAAGGGAAAatcatg…

HMRA1 PCR product generated from T-branch (Fig. 6A, lane 3):
…catgatGTATGTTTTCATTTCAAGGATAGCCTTTGAATCAATT**TAC--AlGT**ATGTAAT
ATGAGAATCAAACTTAAATATATCCTATACTAACActgata…



Extended Data Figure 9. Characterization of circRNAs.

(a) Sanger sequencing confirmed that the PCR products in Figure 5A were derived from T-branches and circRNAs of EFM5 and HMRA1. "/" shows where two ends of exon 2 are ligated. "l" shows where the 5' ss of intron 2 is ligated to the BP of intron 1. The 5' ss and BPS are shown in bold. The BPS contains deletions (show as -) due to errors caused by reverse transcriptase reading through the branch. **(b)** RT-PCR was carried out on RNA extracted from WT yeast cells with or without RNaseR treatment using primers indicated in the schematic diagrams below the gel, indicating that RNase R treatment eliminates linear RNAs. This experiment was repeated four additional times with similar results. **(c)** Protein

and RNA components of E complex assembled on EFM5 IEI-101-M3 pre-mRNA. **(d)** RT-PCR of RNA extracted from BY4742 yeast strain carrying indicated HRMA1 plasmids, with or without RNaseR treatment, using primers shown in the schematic diagrams below the gel. Numbers 246 and 62 designate exon lengths. Lanes 1-3 indicate all constructs were transcribed (endogenous HMRA1 pre-mRNA level is too low to be detected as indicated in lane 3). The HMRA1 middle exon was slightly modified to create a circRNA primer binding site so that only the modified exogenous (*e.g.*, IEI-246 in lane 5) but not WT HMRA1 circRNA (IEI-246 WT in lane 4) can be detected. **(e)** IEI-246-M3 (3xMS2 at the 3' end) RNA or E complex assembled on IEI-246-M3 was incubated with WT or U1-depleted yeast extract in the absence or presence of 30-fold excess competing IEI-246 WT RNA. CircRNA products were monitored using RT-PCR the same way as (d). Experiments in (c) - (e) were repeated one additional time with similar results.

Extended Data Table 1.

Cryo-EM data collection, refinement and validation statistics.

	Ubc4 complex (EMD-0360) (PDB 6N7P)	**Act1 complex (EMD-0361) (PDB 6N7R)**
Data collection and processing		
Magnification	105,000	105,000
Voltage (kV)	300	300
Electron exposure (e^{-}/Å^2)	34.6	29.4
Defocus range (μm)	−1.5 ~ −3.0	−1.5 ~ −3.0
Pixel size (Å)	1.36	1.36
Symmetry imposed	C1	C1
Initial particle images (no.)	1,924,710	3,589,121
Final particle images (no.)	124,825	270,587
Map resolution (Å)	3.6	3.2
FSC threshold	0.143	0.143
Map resolution range (Å)		
Core	3.0-4.5	3.0-4.5
Pre-mRNA helix	---	3.0-6.5
Prp40	15-20	---
Nam8	15-20	---
NCBPs	15-25	---
U1 snRNA	6-15	6-15
Refinement		
Initial model used (PDB code)	n/a	n/a
Model resolution (Å)	4.6	4.3
FSC threshold	0.5	0.5
Model resolution range (Å)	4.6	4.3
Map sharpening *B* factor (Å^2)	−147.1	−94.0
Model composition		
Non-hydrogen atoms	41487	35784

	Ubc4 complex (EMD-0360) (PDB 6N7P)	Act1 complex (EMD-0361) (PDB 6N7R)
Protein residues	4839	3568
Ligands	3	1
B factors (Å^2)		
Protein	57.5	59.9
Ligand	85.3	79.6
R.m.s. deviations		
Bond lengths (Å)	0.01	0.02
Bond angles (°)	1.35	1.50
Validation		
MolProbity score	1.89	1.99
Clashscore	5.09	5.07
Poor rotamers (%)	1.50	2.72
Ramachandran plot		
Favored (%)	92.05	94.21
Allowed (%)	7.35	5.37
Disallowed (%)	0.60	0.42

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by NIH grants GM126157 and GM130673 (R.Z.); GM071940 and AI094386 (Z.H.Z.); and GM122579, GM121487, and CA219847 (R.D.). S.E. is a Howard Hughes Medical Institute Gilliam Fellow. K.K. was supported by an NSF GRFP award and a Stanford Graduate Fellowship. We acknowledge the use of instruments at the Electron Imaging Center for Nanomachines (supported by UCLA and by grants from the NIH (1S10OD018111, 1U24GM116792) and NSF (DBI-1338135 and DMR-1548924)) as well as the CU Anschutz School of Medicine Cryo-EM and proteomics core facilities (partially supported by the School of Medicine and the University of Colorado Cancer Center Support Grant P30CA046934). Molecular graphics and analyses were performed with the UCSF Chimera and ChimeraX, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from NIGMS P41-GM103311 (Chimera, ChimeraX) and NIH R01-GM129325 (ChimeraX). We also thank M. Ares, D. Black, and D. Brow for comments on early versions of the manuscript.

References

1. Zhang L, Vielle A, Espinosa S & Zhao R RNAs in the spliceosome: Insight from cryoEM structures. Wiley interdisciplinary reviews. RNA 10, e1523, doi:10.1002/wrna.1523 (2019). [PubMed: 30729694]
2. Wan R, Bai R, Yan C, Lei J & Shi Y Structures of the Catalytically Activated Yeast Spliceosome Reveal the Mechanism of Branching. Cell, doi:10.1016/j.cell.2019.02.006 (2019).
3. De Conti L, Baralle M & Buratti E Exon and intron definition in pre-mRNA splicing. Wiley interdisciplinary reviews. RNA 4, 49–60, doi:10.1002/wrna.1140 (2013). [PubMed: 23044818]
4. Berget SM Exon recognition in vertebrate splicing. J Biol Chem 270, 2411–2414 (1995). [PubMed: 7852296]

5. Sharma S, Kohlstaedt LA, Damianov A, Rio DC & Black DL Polypyrimidine tract binding protein controls the transition from exon definition to an intron defined spliceosome. Nat Struct Mol Biol 15, 183–191, doi:10.1038/nsmb.1375 (2008). [PubMed: 18193060]
6. Schneider M et al. Exon definition complexes contain the tri-snRNP and can be directly converted into B-like precatalytic splicing complexes. Mol Cell 38, 223–235, doi:10.1016/j.molcel.2010.02.027 (2010). [PubMed: 20417601]
7. Wang PL et al. Circular RNA is expressed across the eukaryotic tree of life. PLoS One 9, e90859, doi:10.1371/journal.pone.0090859 (2014). [PubMed: 24609083]
8. Wilusz JEA 360 degrees view of circular RNAs: From biogenesis to functions. Wiley interdisciplinary reviews. RNA 9, e1478, doi:10.1002/wrna.1478 (2018). [PubMed: 29655315]
9. Starke S et al. Exon circularization requires canonical splice signals. Cell Rep 10, 103–111, doi:10.1016/j.celrep.2014.12.002 (2015). [PubMed: 25543144]
10. Seraphin B, Kretzner L & Rosbash M A U1 snRNA:pre-mRNA base pairing interaction is required early in yeast spliceosome assembly but does not uniquely define the 5' cleavage site. EMBO J 7, 2533–2538 (1988). [PubMed: 3056718]
11. Siliciano PG & Guthrie C 5' splice site selection in yeast: genetic alterations in base-pairing with U1 reveal additional requirements. Genes Dev 2, 1258–1267 (1988). [PubMed: 3060402]
12. Ruby SW & Abelson J An early hierarchic role of U1 small nuclear ribonucleoprotein in spliceosome assembly. Science 242, 1028–1035 (1988). [PubMed: 2973660]
13. Abovich N & Rosbash M Cross-intron bridging interactions in the yeast commitment complex are conserved in mammals. Cell 89, 403–412 (1997). [PubMed: 9150140]
14. Plaschka C, Lin PC, Charenton C & Nagai K Prespliceosome structure provides insights into spliceosome assembly and regulation. Nature 559, 419–422, doi:10.1038/s41586-018-0323-8 (2018). [PubMed: 29995849]
15. Bai R, Wan R, Yan C, Lei J & Shi Y Structures of the fully assembled Saccharomyces cerevisiae spliceosome before activation. Science, doi:10.1126/science.aau0325 (2018).
16. Lewis JD, Izaurralde E, Jarmolowski A, McGuigan C & Mattaj IW A nuclear cap-binding complex facilitates association of U1 snRNP with the cap-proximal 5' splice site. Genes Dev 10, 1683–1698 (1996). [PubMed: 8682298]
17. Qiu ZR, Chico L, Chang J, Shuman S & Schwer B Genetic interactions of hypomorphic mutations in the m7G cap-binding pocket of yeast nuclear cap binding complex: an essential role for Cbc2 in meiosis via splicing of MER3 pre-mRNA. RNA 18, 1996–2011, doi:10.1261/rna.033746.112 (2012). [PubMed: 23002122]
18. Puig O, Gottschalk A, Fabrizio P & Seraphin B Interaction of the U1 snRNP with nonconserved intronic sequences affects 5' splice site selection. Genes Dev 13, 569–580 (1999). [PubMed: 10072385]
19. Lesser CF & Guthrie C Mutational analysis of pre-mRNA splicing in Saccharomyces cerevisiae using a sensitive new reporter gene, CUP1. Genetics 133, 851–863 (1993). [PubMed: 8462846]
20. Liu S et al. Structure of the yeast spliceosomal postcatalytic P complex. Science 358, 1278–1283, doi:10.1126/science.aar3462 (2017). [PubMed: 29146870]
21. Lu M et al. Crystal structure of the three tandem FF domains of the transcription elongation regulator CA150. J Mol Biol 393, 397–408, doi:10.1016/j.jmb.2009.07.086 (2009). [PubMed: 19660470]
22. Liu J, Fan S, Lee CJ, Greenleaf AL & Zhou P Specific interaction of the transcription elongation regulator TCERG1 with RNA polymerase II requires simultaneous phosphorylation at Ser2, Ser5, and Ser7 within the carboxyl-terminal domain repeat. J Biol Chem 288, 10890–10901, doi:10.1074/jbc.M113.460238 (2013). [PubMed: 23436654]
23. Li X et al. CryoEM structure of Saccharomyces cerevisiae U1 snRNP offers insight into alternative splicing. Nature communications 8, 1035, doi:10.1038/s41467-017-01241-9 (2017).
24. Gornemann J et al. Cotranscriptional spliceosome assembly and splicing are independent of the Prp40p WW domain. RNA 17, 2119–2129, doi:10.1261/rna.02646811 (2011). [PubMed: 22020974]
25. Ester C & Uetz P The FF domains of yeast U1 snRNP protein Prp40 mediate interactions with Luc7 and Snu71. BMC Biochem 9, 29, doi:10.1186/1471-2091-9-29 (2008). [PubMed: 19014439]

26. Wiesner S, Stier G, Sattler M & Macias MJ Solution structure and ligand recognition of the WW domain pair of the yeast splicing factor Prp40. J Mol Biol 324, 807–822 (2002). [PubMed: 12460579]
27. Jacewicz A, Chico L, Smith P, Schwer B & Shuman S Structural basis for recognition of intron branchpoint RNA by yeast Msl5 and selective effects of interfacial mutations on splicing of yeast pre-mRNAs. RNA 21, 401–414, doi:10.1261/rna.048942.114 (2015). [PubMed: 25587180]
28. Kappel K & Das R Sampling Native-like Structures of RNA-Protein Complexes through Rosetta Folding and Docking. Structure 27, 140–151 e145, doi:10.1016/j.str.2018.10.001 (2019). [PubMed: 30416038]
29. Howe KJ, Kane CM & Ares M Jr. Perturbation of transcription elongation influences the fidelity of internal exon inclusion in Saccharomyces cerevisiae. RNA 9, 993–1006 (2003). [PubMed: 12869710]
30. Campodonico E & Schwer B ATP-dependent remodeling of the spliceosome: intragenic suppressors of release-defective mutants of Saccharomyces cerevisiae Prp22. Genetics 160, 407–415 (2002). [PubMed: 11861548]
31. Liang D et al. The Output of Protein-Coding Genes Shifts to Circular RNAs When the Pre-mRNA Processing Machinery Is Limiting. Mol Cell 68, 940–954 e943, doi:10.1016/j.molcel.2017.10.034 (2017). [PubMed: 29174924]
32. Ragan C, Goodall GJ, Shirokikh NE & Preiss T Insights into the biogenesis and potential functions of exonic circular RNA. Scientific reports 9, 2048, doi:10.1038/s41598-018-37037-0 (2019). [PubMed: 30765711]
33. Liang D & Wilusz JE Short intronic repeat sequences facilitate circular RNA production. Genes Dev 28, 2233–2247, doi:10.1101/gad.251926.114 (2014). [PubMed: 25281217]
34. Jeck WR et al. Circular RNAs are abundant, conserved, and associated with ALU repeats. RNA 19, 141–157, doi:10.1261/rna.035667.112 (2013). [PubMed: 23249747]
35. Mokry M et al. Accurate SNP and mutation detection by targeted custom microarray-based genomic enrichment of short-fragment sequencing libraries. Nucleic Acids Res 38, e116, doi: 10.1093/nar/gkq072 (2010). [PubMed: 20164091]
36. Spingola M, Grate L, Haussler D & Ares M Jr. Genome-wide bioinformatic and molecular analysis of introns in Saccharomyces cerevisiae. Rna 5, 221–234 (1999). [PubMed: 10024174]
37. Li X et al. Comprehensive in vivo RNA-binding site analyses reveal a role of Prp8 in spliceosomal assembly. Nucleic Acids Res 41, 3805–3818, doi:10.1093/nar/gkt062 (2013). [PubMed: 23393194]
38. Abelson J et al. Conformational dynamics of single pre-mRNA molecules during in vitro splicing. Nat Struct Mol Biol 17, 504–512. [PubMed: 20305654]
39. Carragher B et al. Leginon: an automated system for acquisition of images from vitreous ice specimens. J Struct Biol 132, 33–45, doi:10.1006/jsbi.2000.4314 (2000). [PubMed: 11121305]
40. Zheng SQ, Palovcak E, Armache J-P, Cheng Y & Agard DA in bioRxiv (2016).
41. Rohou A & Grigorieff N CTFFIND4: Fast and accurate defocus estimation from electron micrographs. J Struct Biol 192, 216–221, doi:10.1016/j.jsb.2015.08.008 (2015). [PubMed: 26278980]
42. Scheres SH & Chen S Prevention of overfitting in cryo-EM structure determination. Nature methods 9, 853–854, doi:10.1038/nmeth.2115 (2012). [PubMed: 22842542]
43. Chen S et al. High-resolution noise substitution to measure overfitting and validate resolution in 3D structure determination by single particle electron cryomicroscopy. Ultramicroscopy 135, 24–35, doi:10.1016/j.ultramic.2013.06.004 (2013). [PubMed: 23872039]
44. Rosenthal PB & Henderson R Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. J Mol Biol 333, 721–745 (2003). [PubMed: 14568533]
45. Kucukelbir A, Sigworth FJ & Tagare HD Quantifying the local resolution of cryo-EM density maps. Nature methods 11, 63–65, doi:10.1038/nmeth.2727 (2014). [PubMed: 24213166]
46. Pettersen EF et al. UCSF Chimera--a visualization system for exploratory research and analysis. Journal of computational chemistry 25, 1605–1612, doi:10.1002/jcc.20084 (2004). [PubMed: 15264254]

47. Emsley P, Lohkamp B, Scott WG & Cowtan K Features and development of Coot. Acta Crystallogr D Biol Crystallogr 66, 486–501, doi:10.1107/S0907444910007493 (2010). [PubMed: 20383002]

48. Keating KS & Pyle AM RCrane: semi-automated RNA model building. Acta Crystallogr D Biol Crystallogr 68, 985–995, doi:10.1107/S0907444912018549 (2012). [PubMed: 22868764]

49. Chou FC, Sripakdeevong P, Dibrov SM, Hermann T & Das R Correcting pervasive errors in RNA crystallography through enumerative structure prediction. Nature methods 10, 74–76, doi:10.1038/nmeth.2262 (2013). [PubMed: 23202432]

50. Kappel K et al. De novo computational RNA modeling into cryo-EM maps of large ribonucleoprotein complexes. Nature methods 15, 947–954, doi:10.1038/s41592-018-0172-2 (2018). [PubMed: 30377372]

51. Adams PD et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr 66, 213–221, doi:10.1107/S0907444909052925 (2010). [PubMed: 20124702]

52. Chen VB et al. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr 66, 12–21, doi:10.1107/S0907444909042073 (2010). [PubMed: 20057044]

53. Goddard TD et al. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. Protein Sci 27, 14–25, doi:10.1002/pro.3235 (2018). [PubMed: 28710774]

54. Wisniewski JR, Zougman A, Nagaraj N & Mann M Universal sample preparation method for proteome analysis. Nature methods 6, 359–362, doi:10.1038/nmeth.1322 (2009). [PubMed: 19377485]

55. Grimm M, Zimniak T, Kahraman A & Herzog F xVis: a web server for the schematic visualization and interpretation of crosslink-derived spatial restraints. Nucleic Acids Res 43, W362–369, doi:10.1093/nar/gkv463 (2015). [PubMed: 25956653]

56. Seraphin B & Rosbash M Identification of functional U1 snRNA-pre-mRNA complexes committed to spliceosome assembly and splicing. Cell 59, 349–358 (1989). [PubMed: 2529976]

57. Higuchi F et al. Hippocampal MicroRNA-124 Enhances Chronic Stress Resilience in Mice. The Journal of neuroscience : the official journal of the Society for Neuroscience 36, 7253–7267, doi:10.1523/JNEUROSCI.0319-16.2016 (2016). [PubMed: 27383599]

58. Qin D, Huang L, Wlodaver A, Andrade J & Staley JP Sequencing of lariat termini in S. cerevisiae reveals 5’ splice sites, branch points, and novel splicing events. RNA 22, 237–253, doi:10.1261/rna.052829.115 (2016). [PubMed: 26647463]

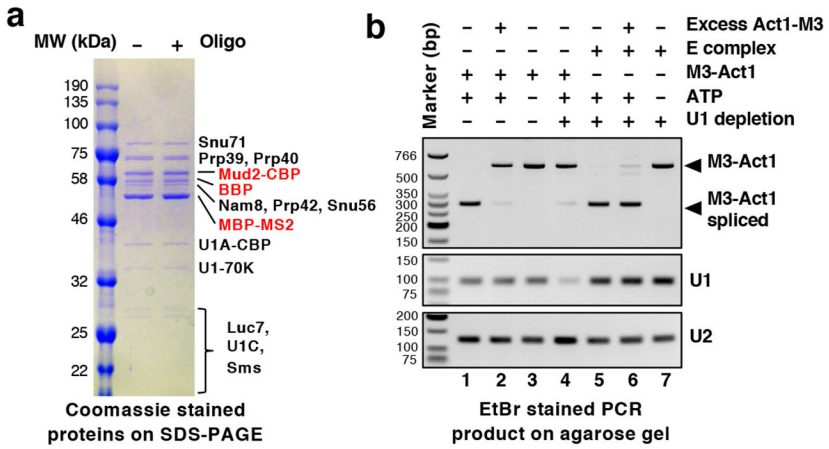


Fig. 1. *In vitro* assembled E complex is functional.

(a) The assembled E complex (with or without DNA oligo-directed RNase H treatment to cleave between the 5' ss and BPS) is purified using the MS2 tag on pre-mRNA and its protein components shown. **(b)** Yeast splicing extract with or without U1 snRNA depletion is incubated with *in vitro* transcribed M3-Act1 or E complex assembled on M3-Act1 in the presence or absence of ATP or excess Act1-M3 (top gel). The splicing outcome is monitored using RT-PCR with primers located in the MS2 binding site region and exon 2 of M3-Act1. The middle and bottom gels demonstrate levels of U1 and U2 snRNA in each sample. Experiments in Fig. 1 were repeated two additional times with similar results. For all gel source data in this paper, see Supplementary Figure 1.

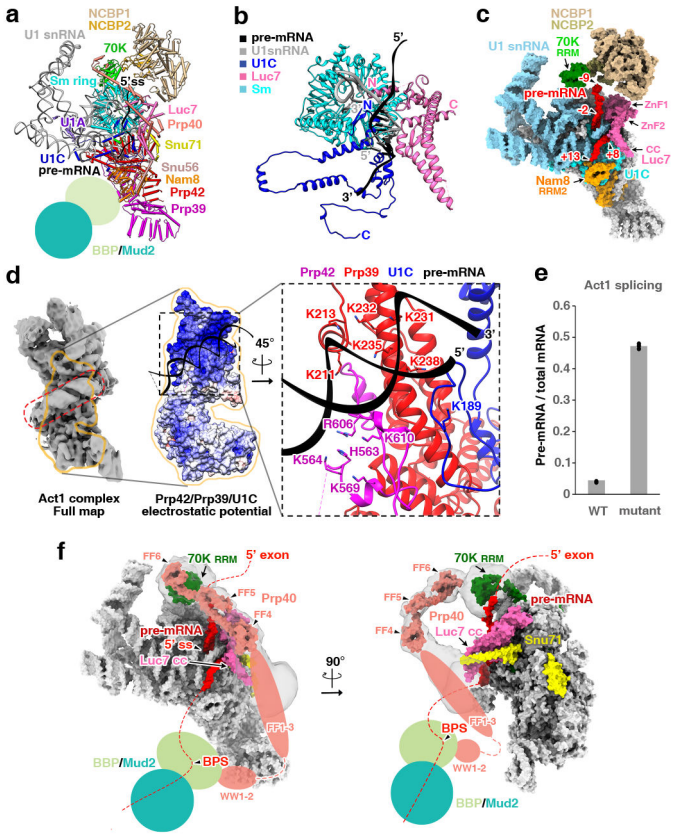


Fig. 2. CryoEM structure of the E complex.

(a) The overall E complex structure. BPP/Mud2 are not modeled due to weak density, but their locations are indicated. **(b)** Ribbon diagrams of protein and RNA models immediately around the 5'ss. **(c)** Surface representation of proteins that are in close proximity to the 5' ss (colored), other proteins (grey), and U1 snRNA (cyan). Pre-mRNA is shown in red and nucleotide positions relative to the 5' ss are labeled (−1 and +1 denote the last nt of the exon and the first nt of the intron, respectively). **(d)** Secondary structure in pre-mRNA. Left: CryoEM density map (filtered to 6 Å) of the entire E complex showing density (in red dashed box) for the pre-mRNA double helix. Middle: Electrostatic potentials of the binding surface for the pre-mRNA double helix. Right: The binding surface formed by Prp39, Prp42, and U1C is shown in ribbon diagrams. Positively charged residues on Prp39 and Prp42 that interact with this double helix are shown in sticks. **(e)** Splicing efficiency of the WT and mutant Act1 intron (that disrupts the secondary structure in the 5' ss to BPS region) in an Act1-Cup1 reporter plasmid, as evaluated by qRT-PCR. Dots represent three technical replicates. This experiment was repeated two additional times with similar results. **(f)** Surface representation of proteins that interact or possibly interact with Prp40 are shown in different colors. Locations of proteins or protein domains not modeled due to weak densities are indicated by various shapes. Transparent grey areas are 8 Å low-pass filtered densities showing likely contacts between Prp40 and U1-70K. Red dashed lines represent hypothetical paths of the pre-mRNA.

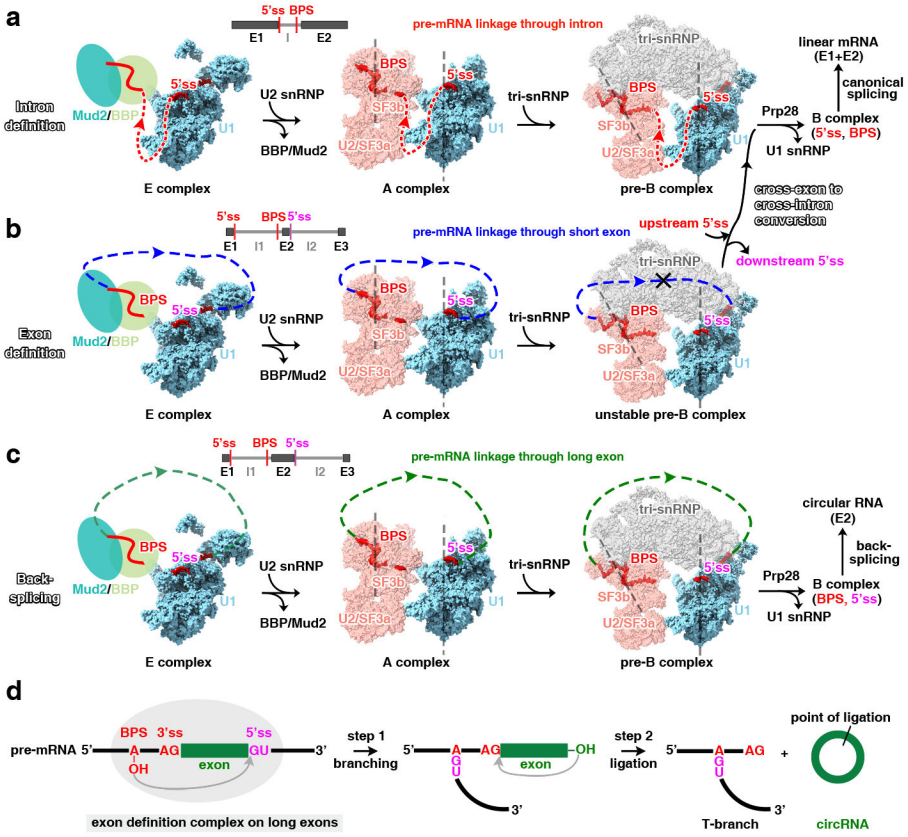


Fig. 3. A unified model for intron definition, exon definition, and back-splicing.

(a) Structures of the E, A, and pre-B complexes are shown in surface representations with U1, U2, and tri-snRNPs in different colors, illustrating the canonical assembly pathway across an intron. Pre-mRNA is shown in red with an arrow indicating the 5' to 3' direction. Red dashed line indicates the hypothetical path of intron connecting the 5' ss and downstream BPS. Vertical dash lines are drawn to denote the orientation of U1 snRNP and U2 SF3b in the A complex. In the pre-B complex, the orientation of U1 snRNP remains the same but that of U2 SF3b is tilted about 30°. **(b)** The same spliceosomal E and A complexes as in (A) can assemble across an exon, but cannot form the pre-B complex on short exons due to steric hindrance. Blue dashed line indicates the hypothetical path of exon connecting the BPS and downstream 5' ss. **(c)** Same as (b), but with a long exon (green dashed line), illustrating that the EDC on long exons can catalyze back-splicing. **(d)** A schematic representation showing how the EDC on a long exon carries out back-splicing and generates circular RNA through the same transesterification reactions used by canonical splicing.

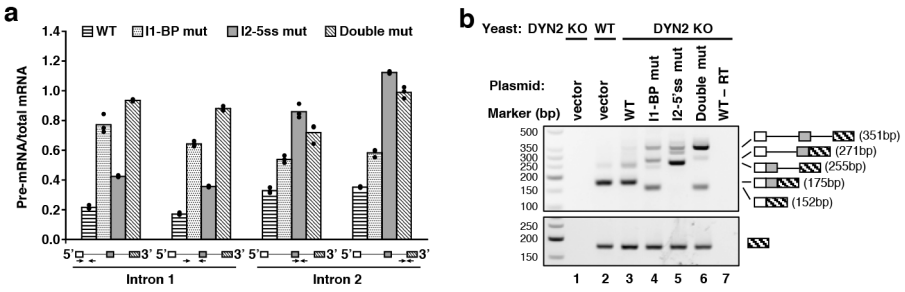


Fig. 4. Exon definition occurs in yeast.

(a) A plasmid containing the WT *DYN2* gene or various mutants was transformed into a *DYN2* KO strain. The splicing efficiency of intron 1 and 2 were evaluated using qRT-PCR with primers specific for intron 1 or intron 2 (indicated by arrows in the schematics under the bar diagram) normalized to total mRNA. Dots represent three technical replicates. **(b)** RT-PCR of RNA extracted from yeast strain carrying indicated plasmids, using primers located in exons 1 and 3 of Dyn2. A schematic of the splicing product and their expected sizes are shown on the right side of the gel. RT-PCR products using primers in exon 3 (bottom gel) serve as an internal quality control of the samples. Experiments in Fig. 4 were repeated two additional times with similar results.

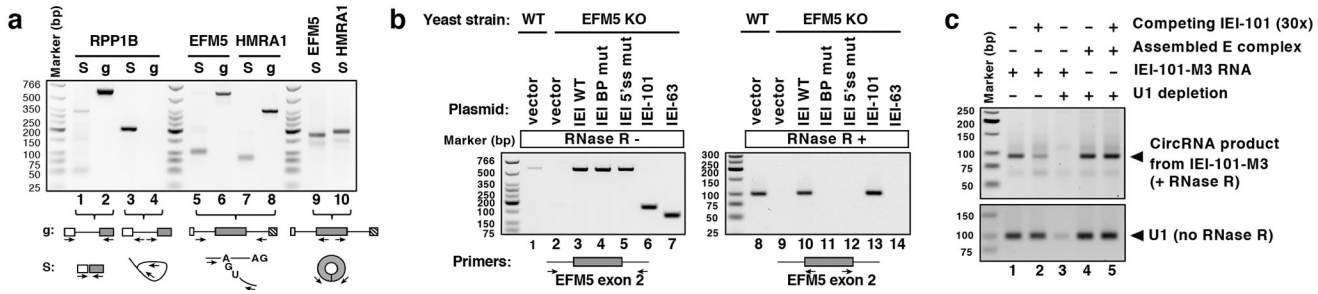


Fig. 5. The EDC catalyzes back-splicing and produces circRNA.

(a) RT-PCR of RNA isolated from spliceosome purified from the Prp22^{H606A} yeast strain (indicated by "S") and PCR using yeast genomic DNA (indicated by "g", as negative controls) for single intron gene RPP1B and multi-intronic genes *EFM5* and *HMRA1* demonstrate the presence of ligated exons (lane 1), lariat (lane 3), T-branches (lanes 5 and 7) and circRNA (lanes 9-10). Primer positions are indicated as arrows in the schematic diagrams below the gel. All images in Fig. 5 are RT-PCR/PCR products on agarose gel with EtBr staining. **(b)** RT-PCR of RNA extracted from WT or *EFM5* KO strain carrying indicated plasmid, with or without RNaseR treatment, using primers shown in the schematic diagrams below the gel. Numbers 101 and 63 designate exon lengths. "mut" represents mutant. Lanes 1-7 indicate all EFM5 constructs are transcribed. **(c)** IEI-101-M3 (3xMS2 at the 3' end) RNA or E complex assembled on IEI-101-M3 was incubated with splicing extract with or without U1 snRNA depletion in the absence or presence of 30-fold excess competing IEI-101 RNA. CircRNA products were monitored the same way as (b). Competing IEI-101 was modified to remove the primer binding sites so it is invisible in the RT-PCR reaction. Experiments in (a), (b), and (c) were repeated one, two, and two additional times, respectively, with similar results.